MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

Final Year Internship Report

on

Data Science Intern at Tiger Analytics

Submitted By:

Yashraj Singh Chouhan

0901CS181124

Faculty Mentor:

Prof. Mir Shahnawaz Ahmad

Asst. Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE

GWALIOR - 474005 (MP) est. 1957

MAY-JUNE 2022

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

Data Science Intern at Tiger Analytics

A final year internship report submitted in partial fulfillment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

Submitted by:

Yashraj Singh Chouhan

0901CS181124

Internship Faculty Mentor:

Prof. Mir Shahnawaz Ahmad

Asst. Professor

Submitted to:

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE

GWALIOR - 474005 (MP) est. 1957

MAY-JUNE 2022

Internship Certificate

Yashraj Singh Chouhan
Jan 18, 2022

Dear Yashraj Singh,

We are pleased to extend to you an offer of internship with **Tiger Analytics India Consulting Private Limited** (the **Company**).

This contract is valid only for the period of internship, and you will be required to sign a separate contract should you take up a full-time role with the Company.

Your internship is subject to the following terms and conditions:

1. Date of Commencement
   The internship is for a period of 4 Months - Jan 24, 2022 to May 31, 2022

2. Place of Work
   Your internship will be administered remotely.

3. Stipend
   You will be paid a stipend of INR 30000 (pre-tax) per month during your internship. This will be deposited into your bank account.

4. Benefits
   Benefits available to full-time employees such as Provident Fund and Medical Insurance are not applicable to Interns.

5. Leave Entitlements
   During your internship period, you are entitled to leave as approved by your manager. Leave cannot be encashed.

6. Safety
   The Company is committed to providing a safe working environment for all employees and therefore required to abide by all safety rules and procedures operating within the Company.

7. Conduct
   You will be expected to dress appropriately for a business setting. Business casual attire as outlined below is considered appropriate:
   a. A collared shirt, pants, and shoes for men
   b. Equivalent Indian or Western business casuals for women

8. Confidentiality
   During your employment with the Company, you will make use of Confidential Information in carrying out your duties. Without limitation, "Confidential Information" includes:
   1. Information relating to the goods and services and proprietary techniques provided by the Company and clients of the Company
   2. All information concerning the business, its methods of operation, marketing and other activities
   3. All databases, lists compiled by the company, client proposals, reports, software, algorithms, and computer programs
   4. Competitive and financial information concerning the business, which is not in the public domain
   5. Information concerning the business of the Company's clients

   You must not, whether during employment or after termination of your employment with the Company, without written authority, divulge 'Confidential Information' to anyone other than an employee authorized to receive the information, or use such information for your own personal gain.

9. Inventions and Copyright
   You assign to the Company your entire right, title and interest in and to any copyright and any industrial or intellectual property rights in any and all works, designs, computer programs, inventions, processes, concepts, strategies, plans and lists (Confidential Property) which (either solely or jointly with others) you have developed or may develop during and/or as a result of your employment with the Company.
   You also agree promptly to disclose to the Company or to its attorneys any and all such Confidential Property developed by you and agree to execute upon demand, at the expense of the Company, all documents which may be desirable to secure to the Company the best copyright, patent or other protection in India and elsewhere and/or rights relating to such Confidential Property.

10. Following End of Internship
    a. Confidentiality
       You agree that upon termination of your internship with the Company you shall return to the Company:
       a. All documents and any other materials constituting or containing Confidential Property or Confidential Information including, without limitation, customers or contacts, correspondence and other written material relating to Confidential Information or Confidential Property and that you will not retain any such documents or material or copies of such documents or material
       b. Company mobile phone or other electronic telecommunications devices that the company has issued to you. The telephone number of the company owned telecommunications devices will remain property of the company

11. Invalidity
    In any terms of provisions in this agreement shall be held illegal or unenforceable, in whole or in part, under any enactment or rule of law, such term or provision or part shall to that extent be deemed not to form part of this agreement but the enforceability of the remainder of this agreement shall not be affected.

12. Variation
    The terms of this contract of employment may be varied by the Company from time to time. You will be notified of any variations.

13. Adherence to Company Policies
    When you join the Company, it will also be a condition of employment that you review and adhere to company policies which you will be notified of subsequently. You agree to adhere to the Company's project financing contracts (e.g. BOT) with the clients.

14. Governing Laws and Jurisdiction
    This contract will be governed by the law in force in Chennai, India.

15. Personal Information and Consent
    By accepting this offer, you are giving your implicit consent to Tiger Analytics to collect and use your personal information for business purposes. Your personal information may be shared with the Clients and prospective Clients of Tiger Analytics as a part of selection or onboarding process to work in projects. Tiger Analytics will also share your personal information with a third party for carrying out the background verification as required. Tiger Analytics will store your employment, financial and personal information during the period of employment and for Data Retention Period after your separation, as per the data retention policy to comply with statutory requirements.

16. Acceptance
    Please sign this letter signifying your acceptance of the appointment and the conditions of service specified in this letter.

    We are pleased to welcome you to the Company. If the preceding terms and conditions of your employment with the Company are acceptable to you, please indicate you acceptance by initialing each page and signing the last page of the attached copy and returning it to me.

Regards
G. Pradeep Kumar
Pradeep Gulipalli
General Manager
Tiger Analytics India LLP

**Since internship is still ongoing, certificate is not available as of now, therefore, as a proof, I've attached snippet of the offer letter for now. However I'll submit the certificate at the time of no dues

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

CERTIFICATE

This is certified that **Yashraj Singh Chouhan (0901CS181124)** has submitted the Internship report titled **Data Science Intern** of the work he has done under the mentorship of **Mir Shahnawaz Ahmad**, in partial fulfillment of the requirement for the award of degree of Bachelor of Technology in Computer Science and Engineering from Madhav Institute of Technology and Science, Gwalior.

Mir Shahnawaz Ahmad
Faculty Mentor
Assistant Professor
Computer Science and Engineering

Dr. Manish Dixit
Professor and Head,
Computer Science and Engineering

Dr. Manish Dixit
Professor & HOD
Department of CSE
M.I.T.S. Gwalior

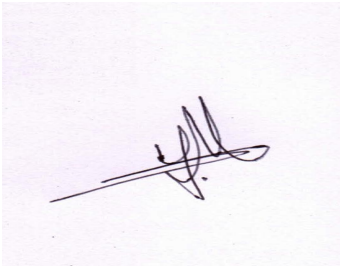
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

DECLARATION

I hereby declare that the work being presented in this Internship report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in CSE at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of Mir Shahnawaz Ahmad, **Asst. Professor**, Department of CSE.

I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.



Yashraj Singh Chouhan

0901CS181124

IV Year,

Computer Science and Engineering

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

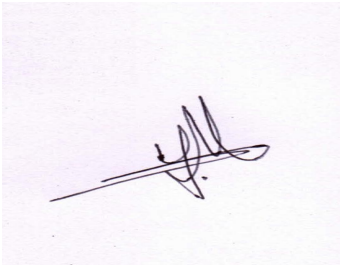
(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

ACKNOWLEDGEMENT

The full semester internship has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary internship as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Department of Computer Science and Engineering, for allowing** me to explore this internship. I humbly thank **Dr. Manish Dixit**, Professor and Head, Department of Computer Science and Engineering, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Mir Shahnawaz Ahmad**, **Asst. Professor**, Department of Computer Science and Engineering, for his continued support and close mentoring throughout the internship. I am also very thankful to the faculty and staff of the department.



Yashraj Singh Chouhan

0901CS181124

IV Year,

Computer Science and Engineering

ABSTRACT

This Data Science Internship at Tiger Analytics, aims at making the candidate job ready by instilling all kinds of necessary skills required to become a successful data analyst. Not just the training programs but a live project mentored by an established personnel in this domain. Data Analytics deals with handling, cleaning, pre-processing, interpreting, analysing, and making meaningful inferences out of the data. This internship commenced on 24-Jan-2022 and is still ongoing. This will finish off after results of 8th semester exams, thus converting my role of an intern into full time Data Analyst at Tiger Analytics.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1: INTRODUCTION

1.1: About Company:

Tiger Analytics is pioneering the ability to do AI and analytics to solve some of the most difficult problems facing companies around the world. We develop customized solutions that leverage data and technology for several Fortune 500 companies. With offices in multiple cities in the United States, United Kingdom, India and Singapore, we have a large number of remote employees around the world. Inc. from recognition as a leader by Forrester Research. And has won multiple awards, including ranking to the fastest growing technology companies by the Financial Times. We are regularly on the well-known list of "Best Analytics Farms". If you're interested in exploring career opportunities at Tiger, we'll give you more details. Do the best job in your business, learn and enjoy a structured approach to innovation.

1.2: About CEO:

Mahesh Kumar is the Founder and CEO of Tiger Analytics. He started Tiger Analytics with a desire to bring his experience in management science to help organizations achieve superior performance through the application of advanced analytics. Before founding Tiger Analytics, Mahesh was on the faculty of the Smith School of Business and Rutgers Business School.

He has conducted research in the areas of data mining and statistical modeling and has successfully applied his research to solve problems related to forecasting, pricing, promotions, and customer segmentation for a wide range of businesses across various verticals. Mahesh holds a Ph.D. in Operations Research and Marketing from MIT, and a B.Tech in Computer Science from IIT Bombay.

1.3: Methodology:

Most data science problems can be divided into three sequential phases – problem definition & data discovery, model estimation & validation, insights & business application. We have broad frameworks to systematically approach a wide variety of data science problems to ensure business value.

1.4: Internship Objectives

- ✓ Internships are generally thought of to be reserved for the college students only, however a wide array of people can be benefited from training.

- ✓ Internships focuses more on training and making a candidate job ready by giving him/her appropriate training in field of specialization along with real time experience by working on business problems.

- ✓ Utilizing internships is a great way to build your resume and learn new skills which are gonna help you in longer period of time making you ready for future endeavors of life, be it corporate job, freelance opportunity or personal project.

- ✓ Internship in order to make candidate aware of real world problems and implementation of solutions outside the textbook.

Currently, we're seeing an IT boom globally, which means a lot of data is over flooding the systems. A lot of this data is irrelevant and complete junk, yet a lot of data is important, carries some values and insights and can be used to make good decisions. This data is needed to be worked upon to get better insights of reality. Such is the base of businesses these days. They generate a lot of data but fail to comprehend and conclude from it. This is one of the application of data science. Computing data and making valuable conclusions from it to allow business to achieve greater heights. Albeit Data Science isn't just restricted to business problems, but in today's commercial world, businesses need this support from Data Science to take the world forward. So this Data Science internship is more specific to business side of the world rather than other applications.

I've been interning at Tiger Analytics as Data Science intern for more than 4 months now. This internship began on 24 Jan' 22. This internship named as Springboard Training Program, where all interns all trained with various modules and made to work on live business problems in order to impart real world knowledge. Modules taught in this program are as follows:

1.5: Modules

1.5.1: Python Programming

Python is one of the most popular programming languages with wide array of applications. One of them is Data Science which makes python one of the baby steps to succeed in this domain. A complete course of python was taught over Udemy in order to achieve fluency in python and familiarity with the application of python in domain of Data Science. The topics taught were:

1: Basics of Python

2: Input/ Output

3: Data types

4: Conditional Statements

5: Loops

6: Basics of OOP

7: Numpy

8: Pandas:

9: Scikit Learn

10: Matplotlib

1.5.2: MS Excel

Importance of MS Excel is widely prevalent in the corporate space. So it does play it's part in Data Science too. MS Excel is used in Data Science for data storage purposes along with data handling and data visualisation. Not that everybody is familiar with technical jargons, so MS Excel comes into play as it is easy to learn and has wide acceptability and support. For someone having basic knowledge of MS Excel is extremely important to survive in the corporate world, weather it be entry level job or GenMan role. The topics taught were:

1: Applications of Office

2: Types of excel sheets

3: Different data types

4: Tables

5: Graphs

6: Data Cleaning

7: Data handling

8: Data Manipulation

9: Data Visualization

10: Formulae implementation

1.5.3: SQL

SQL stands for Structured query Language. It is a relational database language that allows to extract data from tables, a series of selection, sorting and computation criteria, or to update, add or delete new records. In simple words, SQL language is used to do all kinds of data computation by writing queries. SQL plays its role in jobs like data analyst, data engineer, database manager, etc and again, is a vital skill to possess to succeed as a data analyst. Lessons taught were:

1: What is Database

2: Basics of DBMS

3: MySQL workbench

4: Database design

5: Normalization

6: SQL basic commands

7: SQL advanced commands

8: Functions

9: Merging operations

1.5.4: Statistical Concepts

Statistics is backbone of data science. As it is a common knowledge that data science deals with organisation, manipulation, and making inferences out of data, the inferential statistics comes into play. Statistics here is not the regular stats but computations of data using various testing methods, validations, amendments, conclusions and a lot more. Topics taught were:

1: Intro to Statistics

2: Sample vs Population

3: Descriptive Statistics

4: Measures of central tendency

5: Distributions

6: Estimators vs Estimates

7: Confidence Intervals

8: Inferential Statistics

9: Hypothesis Testing

1.5.5: Regression Analysis

Regression is one of the pillars of data Science. Regression is used to Analyse, plan, plot and evaluate various data points through various methods. Lessons taught were:

1: Introduction to Regression

2: Linear Regression

3: Multiple Linear Regression

3: Decision Tree Regression

4: Random Forest Regression

5: Models

6: Evaluation of model performance

1.5.6: GIT and Github

Anyone working in IT sector is well aware of the role and importance of git in a developer's life. Being a necessity, it was taught along wit github, working on linux in order to get wider support at various levels of work.

Chapter 2 : Requirement Analysis

For a project to be successful, it is very important to analyze the project needs as they are collected and throughout the life cycle of the project. Needs analysis helps keep needs in line with business needs. A good needs analysis process will provide a software program that addresses the business objectives set. Requirement Analysis is the process of defining what users expect from an application to be built or modified. Needs analysis involves all activities that are performed to identify the needs of different stakeholders. Needs analysis therefore means analyzing, writing, verifying and managing software or system requirements. The requirements for high quality are documented, feasible, measurable, scalable, traceable, helping to identify business opportunities, and are defined as simplifying system design. To understand, let's have a hypothetical case of classroom. N students having their marks across two subjects A and B.

2.1 Phase 1

Data Cleaning: The whole data has to go through cleaning process, that means removal of junk values, frivolous entries and redundancies. N entries will be checked as per standards and all the issues will be cleared off in order to proceed.

2.2 Phase 2

Data pre-processing: All of the data will be pre-processed in order to make it easy to compute, perform operation upon and calculations. It'll be gone through the standardization to get insights about modelling and gain ideas about approach to solve the questions.

2.3 Phase 3

Data Modelling: Now various questions are dealt upon by formulationg models based on samples and rest population data. These models are then used to get the desired results. Modelling requires implementation of various regression algorithms.

2.4 Phase 4

Statistical Analysis: If statistical analysis is required, it is done in the end to mark finishing of the case.

Chapter 3 : System Requirement Specifications

Following are the system requirements for a data analyst:

Hardware Specifications:

Resource	Requirements	
Operating systems (64-bit)	Microsoft Windows	• Windows Server 2008 R2 • Windows Server 2008 R2 Standard • Windows Server 2008 R2 Enterprise • Windows Datacenter 2008 R2 • Windows Server 2012 • Windows Server 2012 R2 • Windows 2012 R2 Datacenter • Windows Server 2016 • Windows Server 2016 Datacenter • Windows Server 2019
	Linux	• Red Hat Enterprise Linux versions 6.x, 7.x • SUSE Linux Enterprise Server 11.x, 12.x, 15 • Oracle Linux Server 6.x, 7.x
	The preceding list is applicable for the product and the target hosts where data resides that you might install on the target hosts.	
CPU and memory	2 CPUs and 4 GB RAM **Note:** In a production environment, the minimum requirement would be 8 CPUs and 16 GB RAM for handling a 50-100 GB volume of data per day and up to four users.	

Fig 3.1

Software Requirements:

Resource	Requirements
Port	For more information about ports, see Communication ports and protocols.
Web browsers	• Mozilla Firefox (latest) • Google Chrome (latest) • Safari on Mac OS -10.x,11 • Internet Explorer 11.x
Screen resolution (to view the Console)	1280*1024
Java (for running CLIs)	• Oracle JRE 1.8.0, build 152 (bundled with IT Data Analytics 11.3.01) • Azul JRE 1.8.0, build 202 Azul 8 Java is supported on version 11.3.02 and later of the product.

Fig 3.2

Chapter 4 : Technologies used during internship

Data Science comprises of a multitude of technologies. It consists of various roles, which work at various levels, incorporating various technologies, helping world manage their data in an effective and efficient way. Following are the technologies used in Data Science.

Basics:

1: Python Programming

2: SQL

3: MS Excel

4: Git

5: JupyterLab

6: Google Colaboratory

Intermediate:

1: Regression Analysis

2: Statistical Analysis

3: AWS

4: MS Azure

5: Arduino

Advanced:

1: Hadoop

2: Tableau

3: PowerBi

4: Machine Learning

5: Artificial Intelligence

6: TensorFlow

Chapter 5 : Project (Work done during internship)

The Project was a business case of a company which gave us various datasets in order to compute the data and fulfill all the deliverables. Although the data can't be shared due to **non - disclosure agreement**. Still a brief summary of the work done is present in this file.

5.1 Data Processing

Data processing involves converting raw data into an understandable format. Improving data efficiency is an important part of data mining. This method directly affects the results of the analysis algorithm. Preprocessing data is typically done in six simple steps.

5.1.1: Gathering the data:

Data is raw information and represents observations of both human and machine worlds. Recording depends entirely on the type of problem you want to solve. Each machine learning problem has its own approach.

5.1.2: Import the dataset & Libraries

The first step is usually to import the libraries needed for your program. A library is basically a collection of modules that you can access and use. You can also use the "import" keyword to import the library into your Python code.

5.1.3: Dealing with Missing Values

Sometimes some data is missing from the dataset. Once found, you can remove these rows or calculate the mean, mode, or median of the features and replace them with the missing values. This is an approximation that allows you to add variance to your dataset.

5.1.4: Divide the dataset into Dependent & Independent variable

After importing the dataset, the next step is to identify the independent variable (X) and the dependent variable (Y).

5.1.5. Split the dataset into training and test set

Machine learning typically divides the data into training and test data and applies the model. Typically, you split the dataset into 70:30 or 80:20 (depending on your requirements). This means that 70% of the data will be used for training and 30% of the data will be used for testing. This task imports train_test_split from scikit's model_selection library. Then, to create a training set and a test set, X_train (training some of the features), X_test (testing some of the features), Y_train (some of the dependent variables associated with the X train set). To train) to create four sets. Same index), Y_test (test part of the dependent variable associated with the X test set, and therefore the same index). Assign a train_test_split that takes parameters — an array

(X and Y), test_size (the ideal choice is to allocate 20% of the dataset to the test set, usually assigned as 0.2, where 0.25 means 25 increase).

5.1.6: Feature Scaling

The final step in data preprocessing is to apply very important functional scaling. Feature scaling is a technique for standardizing independent features that exist in data to a fixed range. Executed during data preprocessing. Scaling Reasons: In most cases, datasets contain features that vary greatly in size, unit, and range. However, this is a problem because most machine learning algorithms use the Euclidean distance between two data points in their calculations.

5.2 Inferential Statistics

Inference statistics are a branch of statistics that make inferences about population data from sample data using a variety of analytical tools. In addition to inference statistics, descriptive statistics form another branch of statistics. Inference statistics help draw inferences about the population, and descriptive statistics summarize the characteristics of the dataset.

There are two main types of hypothesis testing in inference statistics and regression analysis. The sample selected in the inference statistics must be representative of the entire population. Inference statistics help you better understand the population data by analyzing the samples taken from the population data. Use a variety of analytical tests and tools to help you make generalizations about the population. Many sampling techniques are used to select random samples that accurately represent the population. Some of the important methods are simple random sampling, stratified sampling, cluster sampling, and systematic sampling techniques. Inference statistics can be defined as areas of statistics that draw inferences about the population by examining random samples using analytical tools. The purpose of inference statistics is to make generalizations about the population. In inference statistics, statistics are population parameters, for example:

5.2.1: Hypothesis Testing

Hypothesis testing is a type of inference statistic used to test assumptions and derive inferences about the population from the available sample data. This includes setting up null and alternative hypotheses and then performing a statistical significance test. Conclusions are drawn based on test statistic values, critical values, and confidence intervals. Hypothesis tests are on the left, right, and both sides. Below are some important hypothesis tests used in inference statistics.

Z-test: The Z-test follows a normal distribution and applies to data with a sample size of 30 or larger. Used to test if the sample mean and the population mean are equal if the population variance is known.

T-test: If the data follow Student's t-distribution and the sample size is less than 30, the t-test is used. Used to

compare a sample to the population mean when the population variance is unknown.

F-test: The F-test is used to test if there is a difference between the variances of two samples or populations.

Confidence Intervals: Confidence intervals are useful for estimating population parameters. For example, a 95% confidence interval indicates that if you run the test 100 times on a new sample under the same conditions, you can expect the estimate to be 95 times within the specified interval. Confidence intervals are also useful in calculating critical values in hypothesis testing.

Apart from these tests, other tests used in inference statistics include the ANOVA test, Wilcoxon signed rank test, Mann-Whitney U test, and Clascal Wallis H test.

Hypothesis Testing

www.FairlyNerdy.com

Type Of Test	Purpose	Example	Equation	Comment	Excel Function
Z Test	Test if the average of a single population is equal to a target value	Do babies born at this hospital weigh more than the city average	$Z = \frac{\bar{x} - u_0}{\frac{\sigma}{\sqrt{n}}}$	Z test does not need df σ = population standard deviation	=Ztest(array,x,sigma)
1 Sample T-Test	Test if the average of a single population is equal to a target value	Is the average height of male college students greater than 6.0 feet?	$t = \frac{\bar{x} - u_0}{\frac{s}{\sqrt{n}}}$ $df = n - 1$	s = sample standard deviation	no built in equation use =STDEVA for standard deviation use =AVERAGE for mean use =T.DIST.RT to get 1 tailed confidence use =T.DIST.2T to get 2 tailed confidence
Paired T-Test	Test if the average of the differences between paired or dependent samples is equal to a target value	Weigh a set of people. Put them on a diet plan. Weigh them after. Is the average weight loss significant enough to conclude the diet works?	$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$ $df = n - 1$	d bar = average difference between samples s = sample deviation of the difference n = count of one set of the pairs (don't double count)	=TTEST(Array1,Array2,*,1) * -> 1 for 1 tailed, 2 for 2 tailed
2 Sample T-Test Equal Variance	Test if the difference between the averages of two independent populations is equal to a target value	Do cats eat more of type A food than type B food	$df = n_1 + n_2 - 2$ $t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	n1, n2 = count of sample 1, 2	=TTEST(Array1,Array2,*,2)
2 Sample T-Test Unequal Variance	Test if the difference between the averages of two independent populations is equal to a target value	Is the average speed of cyclists during rush hour greater than the average speed of drivers	$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$		=TTEST(Array1,Array2,*,3)

Fig 5.1

5.2.2: Regression Analysis

Regression analysis is a statistical technique used to determine the structure of relationships between two variables (simple linear regression) or three or more variables (multiple regression). According to the Harvard Business School Business Analysis Online Course, Regression is used for two main purposes.

1. To investigate the size and structure of relationships between variables
2. To predict variables based on relationships with other variables

Both of these insights can influence strategic business decisions. there is.
"Regression gives us insight into the structure of the relationship and how well the data fits into the relationship," says Professor HBS, who teaches business analysis, one of three courses that teach qualifications for preparation. Said Jan Hammond. CORe) Program. “We find that such insights are very useful for analyzing past trends and making predictions.” One way to think of
regression is to use an independent variable on the x-axis and a dependent variable on the Y-axis. Is to create a scatter plot of the data. shaft. Regression lines are the best lines for scatter plot data. The regression equation plots the slope of the line and the relationship between the two variables, along with an estimate of the error.
Physically creating this scatter plot can provide a natural starting point for analyzing relationships between variables.

Types of Regression Analysis

There are two types of regression analysis: single variable linear regression and multiple regression.

Single variable linear regression is used to determine the relationship between two variables: the independent and dependent. The equation for a single variable linear regression looks like this:

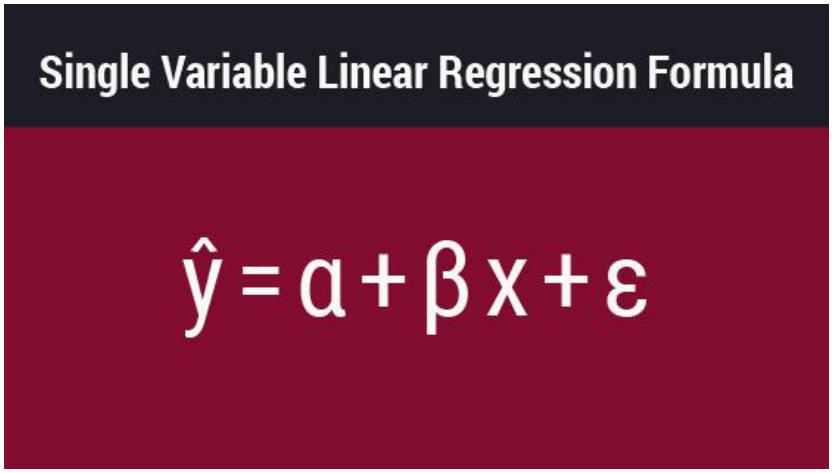


Fig 5.2

In the equation:

- **ŷ** is the expected value of Y (the dependent variable) for a given value of X (the independent variable).
- **x** is the independent variable.

- **α** is the Y-intercept, the point at which the regression line intersects with the vertical axis.
- **β** is the slope of the regression line, or the average change in the dependent variable as the independent variable increases by one.
- **ε** is the error term, equal to Y – ŷ, or the difference between the actual value of the dependent variable and its expected value.

Multiple regression, on the other hand, is used to determine the relationship between three or more variables: the dependent variable and at least two independent variables. The multiple regression equation looks complex but is similar to the single variable linear regression equation:

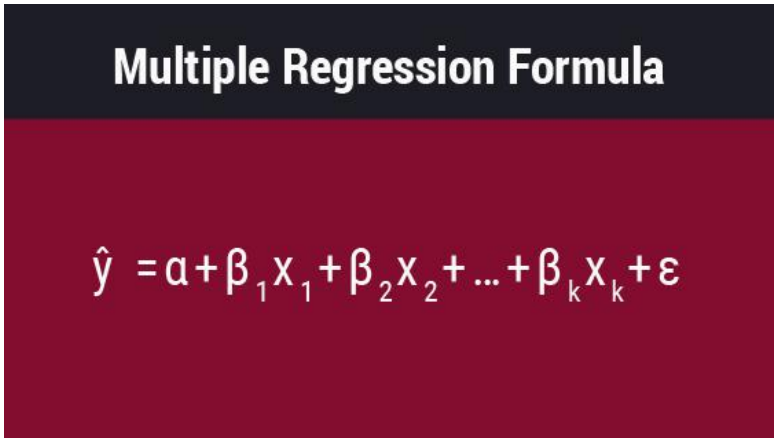


Fig 5.3

Each component of this equation is represented in the same way as the previous equation, with the subscript k added. It represents the total number of independent variables that can be examined. For each independent variable to include in the regression, multiply the slope of the regression line by the value of the independent variable and add it to the rest of the equation.

How to perform regressions You can perform both single-variable linear regressions and multiple regressions using various statistical programs such as

Microsoft Excel, SPSS, and STATA. If you are interested in hands-on practice with this skill, Business Analytics can create scatter plots in Microsoft Excel to perform regressions, understand the output, and use it for business intent. Teach learners how to drive decisions.

Confidence calculation and error description :

It is important to note. This regression analysis overview is introductory and does not cover confidence levels, significance, variance, and error calculations. If you are working with a statistical program, these

calculations may be provided or you may need to implement a function. When performing regression analysis, these metrics are important for assessing how important and how important the results are.

5.3 Statistical Plots

Various statistical plots are used for data visualization. Some of graphs used are:

5.3.1 Bar Graph

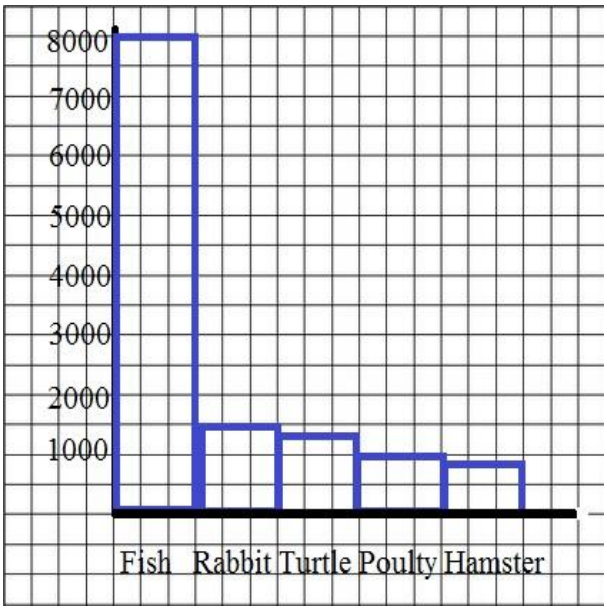


Fig 5.4

5.3.2: Segmented Bar Graph

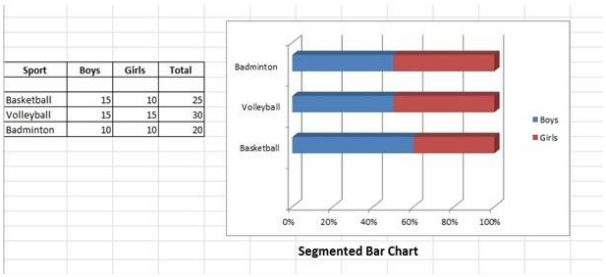


Fig 5.5

5.3.3. Box and Whiskers (Boxplots)

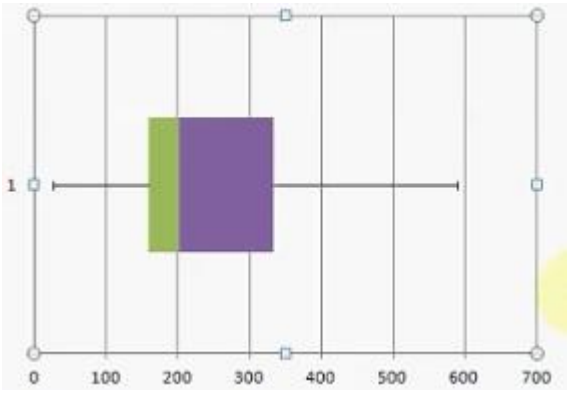


Fig 5.6

5.3.4: Frequency Polygon

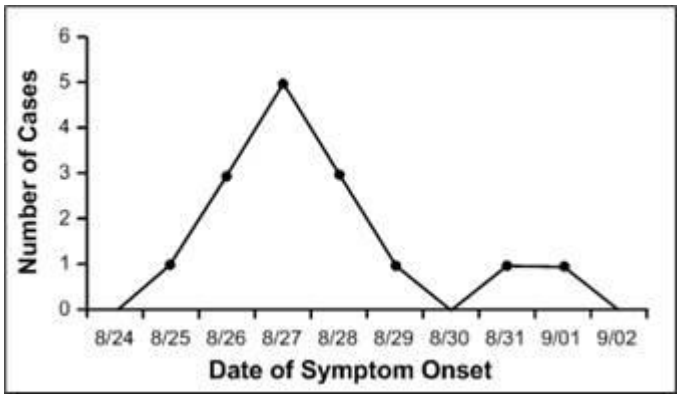


Fig 5.7

5.3.5. Histogram

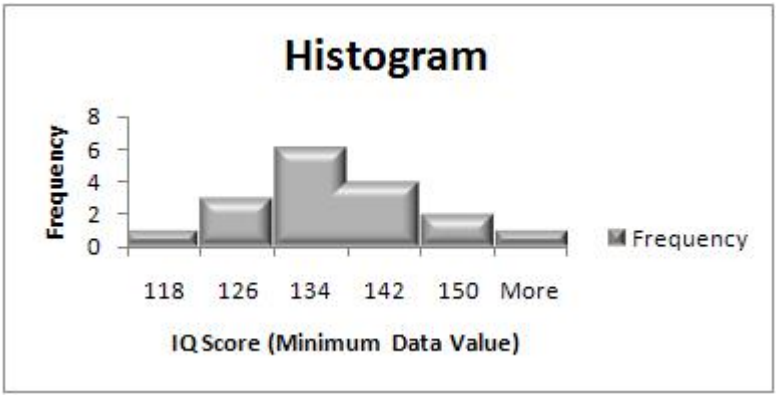


Fig 5.8

5.3.6: Pie Graphs

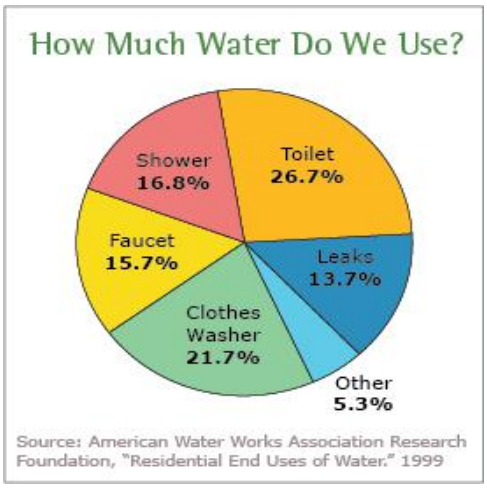


Fig 5.9

5.3.7. Scatter Graphs

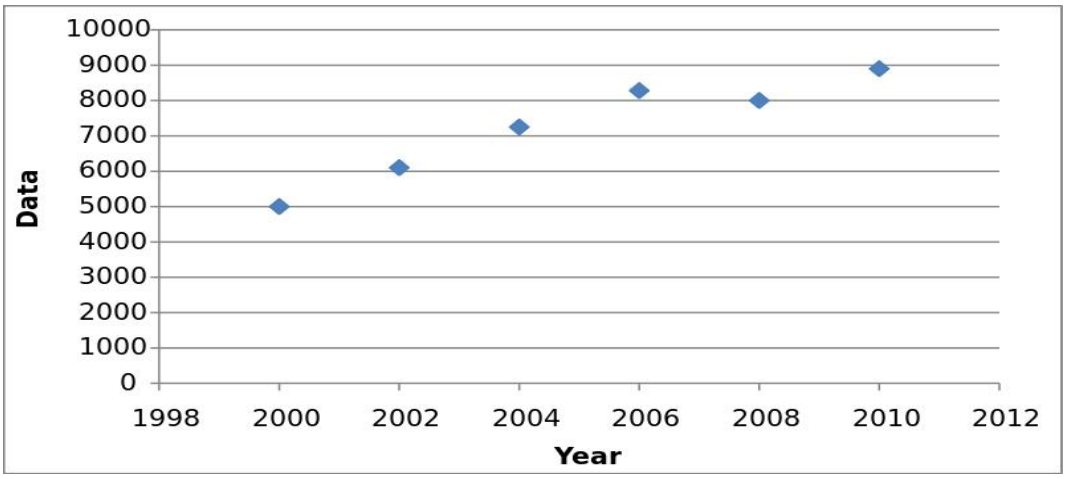


Fig 5.10

Chapter 6 : Final Analysis and Results

6.1: Problems faced:

Though it was a smooth internship, there were few problems that I faced. Some of them are:

- I had 0 prior experience to work with git. So getting along with git was a bit of challenge.
- This is the first time I've worked for a corporate entity, so it took me little time to get going with the processes.
- UBUNTU was completely alien to me so getting along with that was again a problem yet conquered easily with time.
- Using the standard code templates of Tiger Analytics was a bit of hassle as it was based on Ubuntu, and as I've mentioned already, Ubuntu did cause a little problem.

6.2: Limitations:

Though it was pretty independent and self dependent, there were a few limitations. Some of them are:

- We were confined to modules of data analytics only. No exposure to big data was given.
- Working on ubuntu caused some limitations, as it doesn't get windows like universal support.

Conclusion

In the end, I can say I've gained a lot of knowledge regarding the domain of data analytics comprising of all modules and I'm now ready to step in the shoes of full time Data Analyst at Tiger Analytics. Since it's still ongoing internship, I guess it's too early to call a result but I am confident with my skills and grateful to MITS and Tiger Analytics for this wonderful opportunity.

References

1: Udemy - Data Analytics courseware

2: Tiger Analytics - Conpany Information

3: Byjus - Statistical Plots

4: Tiger Academy - Course structure and SpringBoard Training

FPR:

FPR 1

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	Yashraj Singh Chouhan		Department	Data Science Core	
Industry/Organization	Tiger Analytics		Date/Duration	24 jan - 11 feb	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					
Learning capacity/Knowledge up gradation					
Performance/Quality of work					
Behaviour/Discipline/Team work					
Sincerity/Hard work					
Comment on nature of work done/Area/Topic	Yashraj Singh Chouhan has started with the campus Training Program & started going through all the fundamental modules of data science. He is completing courses on time & performs well.				
OVERALL GRADE (Any one)	**VERY GOOD**				
Name of Industry Mentor	Padmajothi Murugaboopathy				
Signature of Industry Mentor	Padmajothi Murugaboopathy				

Receiving Date		**Name of Faculty Mentor**		**Sign**	

FPR 2

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	Yashraj Singh Chouhan		Department	Data Science Core	
Industry/Organization	Tiger Analytics		Date/Duration	14 feb – 1 mar	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					
Learning capacity/Knowledge up gradation					
Performance/Quality of work					
Behaviour/Discipline/Team work					
Sincerity/Hard work					
Comment on nature of work done/Area/Topic	Yashraj Singh Chouhan has finished the foundations module and has moved on to the basics of statistics.				
OVERALL GRADE (Any one)	**Excellent**				
Name of Industry Mentor	Padmajothi Murugaboopathy				
Signature of Industry Mentor	Padmajothi Murugaboopathy				

Receiving Date	02/03/22	**Name of Faculty Mentor**	Mir Shahnawaz Ahmad	**Sign**	

FPR 3

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	Yashraj Singh Chouhan		Department	Data Science Core	
Industry/Organization	Tiger Analytics		Date/Duration	1 Mar – 16 Mar	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					
Learning capacity/Knowledge up gradation					
Performance/Quality of work					
Behaviour/Discipline/Team work					
Sincerity/Hard work					
Comment on nature of work done/Area/Topic	Yashraj Singh Chouhan is doing well with current modules of MLE, git and Regression analysis with timely assignment submissions.				
OVERALL GRADE (Any one)	**Excellent**				
Name of Industry Mentor	Padmajothi Murugaboopathy				
Signature of Industry Mentor	Padmajothi Murugaboopathy				

Receiving Date	16/03/22	**Name of Faculty Mentor**	Mir Shahnawaz Ahmad	**Sign**	

FPR 4

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	Yashraj Singh Chouhan		Department	Data Science Core	
Industry/Organization	Tiger Analytics		Date/Duration	17 Mar – 31 Mar	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					
Learning capacity/Knowledge up gradation					
Performance/Quality of work					
Behaviour/Discipline/Team work					
Sincerity/Hard work					
Comment on nature of work done/Area/Topic	Yashraj Singh Chouhan is doing well with current modules of Regression analysis and Case studies with timely assignment submissions.				
OVERALL GRADE (Any one)	**Excellent**				
Name of Industry Mentor	Padmajothi Murugaboopathy				
Signature of Industry Mentor	Padmajothi Murugaboopathy				

Receiving Date	30/03/2022	**Name of Faculty Mentor**	Mr. Mir Shahnawaz Ahmad	**Sign**	

FPR 5

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	Yashraj Singh Chouhan		Department	Data Science Core	
Industry/Organization	Tiger Analytics		Date/Duration	20 Mar – 12 Apr	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work				■	
Learning capacity/Knowledge up gradation					■
Performance/Quality of work					■
Behaviour/Discipline/Team work					■
Sincerity/Hard work					■
Comment on nature of work done/Area/Topic	Yashraj Singh Chouhan has finished with basic modules of ML.				
OVERALL GRADE (Any one)	**Very Good**				
Name of Industry Mentor	Padmajothi Murugaboopathy				
Signature of Industry Mentor	Padmajothi Murugaboopathy				

Receiving Date		**Name of Faculty Mentor**		**Sign**	

FPR 6

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	Yashraj Singh Chouhan		Department	Data Science Core	
Industry/Organization	Tiger Analytics		Date/Duration	15 Mar – 29 Apr	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work				■	
Learning capacity/Knowledge up gradation					■
Performance/Quality of work					■
Behaviour/Discipline/Team work					■
Sincerity/Hard work					■
Comment on nature of work done/Area/Topic	Yashraj Singh Chouhan has started working on template implementation and collaborative workflow.				
OVERALL GRADE (Any one)	**Very Good**				
Name of Industry Mentor	Padmajothi Murugaboopathy				
Signature of Industry Mentor	Padmajothi Murugaboopathy				

Receiving Date		**Name of Faculty Mentor**		**Sign**	

FPR 7

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	Yashraj Singh Chouhan		Department	Data Science Core	
Industry/Organization	Tiger Analytics		Date/Duration	30 Apr – 15 May	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					
Learning capacity/Knowledge up gradation					
Performance/Quality of work					
Behaviour/Discipline/Team work					
Sincerity/Hard work					
Comment on nature of work done/Area/Topic	Yashraj Singh Chouhan is working on business case, project followed by timely submissions.				
OVERALL GRADE (Any one)	**Very Good**				
Name of Industry Mentor	Padmajothi Murugaboopathy				
Signature of Industry Mentor	Padmajothi Murugaboopathy				

Receiving Date	15/05/2022	**Name of Faculty Mentor**	Mir Shahnawaz Ahmad	**Sign**	