IBM DATA ANALYST

Summer Internship Report Submitted in partial

fulfilment of the requirement for undergraduate degree

of

Bachelors of Technology

In

Computer Science and Engineering

By

Parvathaneni Kaushik

221810308036

Under the guidance of

Mrs. K Vani Prasanna

Department of Electronics and Communication Engineering
GITAM School of Technology
GITAM (Deemed to be University)
Hyderabad - 502329

November 2021

DECLARATION

I submit this industrial training work entitled **"IBM DATA ANALYST CAPSTONE PROJECT**" to GITAM (Deemed TO Be University), Hyderabad in partial fulfilment of the requirements for the award of the degree of "**Bachelor of Technology"** in "**Computer Science and Engineering"**. I declare that it was carried out independently by me under the guidance of Mrs. K Vani Prasanna, Asst. Professor, GITAM (Deemed to Be University), Hyderabad, India.

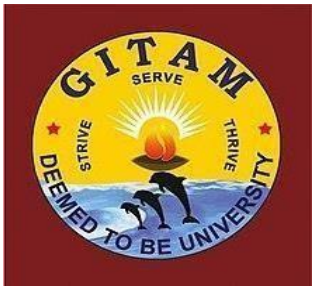
The results embodied in this report have not been submitted to any other University or Institute for the award of any degree or diploma.

Place: HYDERABAD

Student Name: P.Kaushik

Date:

Student Roll No:221810308036



Department of Computer Science and Engineering GITAM (Deemed to be University)
Rudraram Mandal, Sangareddy district, Patancheruvu, Hyderabad, Telangana 502329

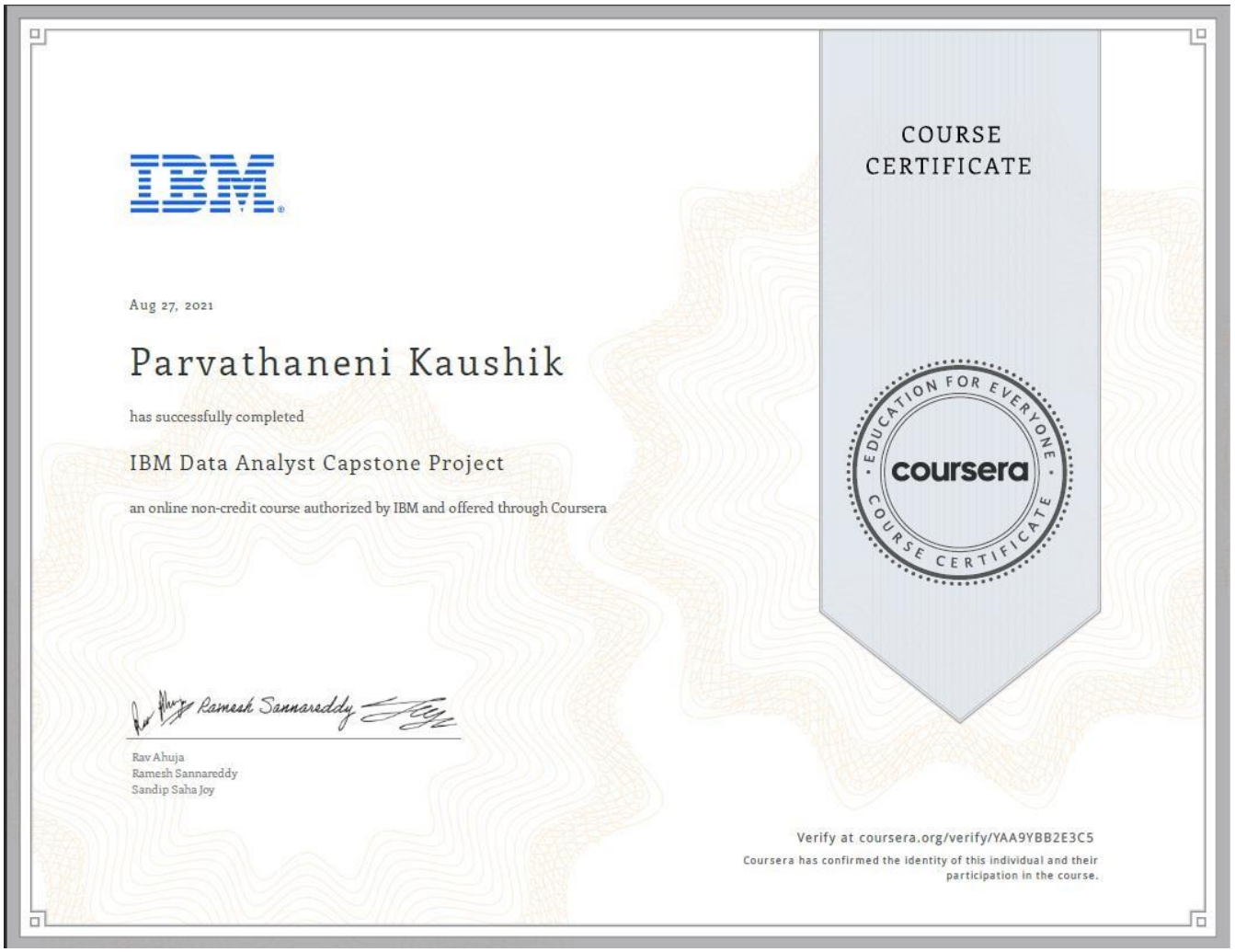
CERTIFICATE

This is to certify that the Industrial Training Report entitled **"IBM DATA ANALYST CAPSTONE PROJECT"** is being submitted by P.Kaushik (221810308036) in partial fulfilment of the requirement for the award of **Bachelor of Technology in Computer Science and Engineering** at GITAM (Deemed to Be University), Hyderabad during the academic year 2021-2022.

It is faithful record work carried out by him at the **Computer Science Engineering Department**, GITAM School of Technology, GITAM Deemed to be University, Hyderabad Campus under my guidance and supervision.

Mrs. K Vani Prasanna
Assistant Professor
Department of CSE

Prof. S. Phani Kumar
Professor and HOD
Department of CSE



COURSE CERTIFICATE

Aug 27, 2021

Parvathaneni Kaushik

has successfully completed

IBM Data Analyst Capstone Project

an online non-credit course authorized by IBM and offered through Coursera

Rav Ahuja
Ramesh Sannareddy
Sandip Saha Joy

Verify at coursera.org/verify/YAA9YBB2E3C5

Coursera has confirmed the identity of this individual and their participation in the course.

ACKNOWLEDGEMENT

This internship would not have been successful without the help of several people. I would like to thank the personalities who were part of project in numerous ways, those who gave us outstanding support from the birth of the seminar.

I am extremely thankful to our honorable Pro-Vice Chancellor, Prof. Siva Prasad for providing necessary infrastructure and resources for the accomplishment of this seminar.

I am very much obliged to our beloved Prof.S.Phani Kumar, Head of the Department of Computer Science & Engineering for providing the opportunity to undertake this project and encouragement in completion of this seminar.

I hereby wish to express a deep sense of gratitude to Mrs. K Vani Prasanna , Assistant Professor, Department of Computer Science and Engineering, for their esteemed guidance, moral support and invaluable advice provided by them for the success of this internship.

I am also thankful to all the staff members of Computer Science and Engineering department who have cooperated in making this project a success. I would like to thank my parents and friends who extended their help, encouragement and moral support either directly or indirectly in this seminar.

Sincerely,

P.Kaushik

ABSTRACT

In this course we will apply various Data Analytics skills and techniques that we have learned as part of the previous courses in the IBM Data Analyst Professional Certificate. we will assume the role of an Associate Data Analyst who has recently joined the organization and be presented with a business challenge that requires data analysis to be performed on real-world datasets.

We will undertake the tasks of collecting data from multiple sources, performing exploratory data analysis, data wrangling and preparation, statistical analysis and mining the data, creating charts and plots to visualize data, and building an interactive dashboard. The project will culminate with a presentation of your data analysis report, with an executive summary for the various stakeholders in the organization.

This project is a great opportunity to showcase Data Analytics skills, and demonstrate proficiency to potential employers.

CHAPTER I: INTRODUCTION

CHAPTER II: DATA ANALYSIS

CHAPTER III: DATA COLLECTION

CHAPTER IV: DATA WRANGLING

CHAPTER V: EXPLORATARY DATA ANALYSIS

CHAPTER VI: DATA VISUALIZATION

CHAPTER VII: DASHBOARD

CHAPTER VIII: CONCLUSION

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION TO CAPSTONE PROJECT

You have recently been hired as a Data Analyst by a global IT and business consulting services firm that is known for their expertise in IT solutions and their team of highly experienced IT consultants. In order to keep pace with changing technologies and remain competitive, your organization regularly analyzes data to help identify future skill requirements.

As a Data Analyst, you will be assisting with this initiative and have been tasked with collecting data from various sources and identifying trends for this year's report on emerging skills.

Your first task is to collect the top programming skills that are most in demand from various sources including: Job postings
Training portals
Surveys

Once you have collected enough data, you will begin analyzing the data and identify insights and trends that may include the following:

What are the top programming languages in demand?

What are the top database skills in demand?

What are the popular IDEs?

You will begin by scraping internet web sites and accessing APIs to collect data in various formats like .csv files, excel sheets, and databases.
Once this is completed, you will make that data ready for analysis using data wrangling techniques. When the data is ready you will then want to apply statistical techniques to analyze the data. Then bring all of your information together by using IBM Cognos Analytics to create your dashboard. And finally, show off your storytelling skills by sharing your findings in a presentation.

You will be evaluated using quizzes in each module as well as the final project presentation.

CHAPTER 2 DATA ANALYSIS

2.1 INTRODUCTION TO DATA ANATYSIS.

Data analysis is an internal organizational function performed by Data Analysts that is more than merely presenting numbers and figures to management. It requires a much more in-depth approach to recording, analyzing and dissecting data, and presenting the findings in an easilydigestible format. With a data analysis you'll be able to provide a company with decisionmaking insight into the following key areas:

- Predict customer trends and behaviors
- Analyse, interpret and deliver data in meaningful ways
- Increase business productivity
- Drive effective decision-making

Data analysis is important in business to understand problems facing an organization, and to explore data in meaningful ways. Data in itself is merely facts and figures. Data analysis organizes, interprets, structures and presents the data into useful information that provides context for the data. This context can then be used by decision-makers to take action with the aim of enhancing productivity and business gain.

Their responsibilities around analyzing data help the business managers make informed decisions to drive the company forward, improve efficiency, increase profits and achieve organizational goals.

To do this effectively, Data Analysts need to be able to:

- Understand business direction and objectives
- Explore the meaning behind the numbers and figures in data
- Analyse the causes of certain events based on data findings
- Present technical insights using easy-to-understand language
- Contribute to business decision-making by offering educated opinions

2.2 Types and Principles of Data Analysis:

2.2.1 Types of Data:

Defining data is that data is numbers, characters, images, or other method of recording, in a form which can be assessed to make a determination or decision about a specific action. Many believe that data on its own has no meaning, only when interpreted does it take on meaning and become information. By closely examining data we can find patterns to perceive information, and then information can be used to enhance knowledge.

Qualitative data:

Data that is represented either in a verbal or narrative format is qualitative data. These types of data are collected through focus groups, interviews, opened ended questionnaire items, and other less structured situations. A simple way to look at qualitative data is to think of qualitative data in the form of words.

Quantitative data:

Quantitative data is data that is expressed in numerical terms, in which the numeric values could be large or small. Numerical values may correspond to a specific category or label.

2.2.2 Principles for Data Analysis:

1. To provide information to program staff from a variety of different backgrounds and levels of prior experience.
2. To create a “value-added” framework that presents strategies, concepts, procedures, methods and techniques in the context of real-life examples.
3. To appreciate that learning takes time.
4. Comfort, confidence, and competence take practice.
5. Data analysis provides opportunities to “reduce the burden.”

2.3 DATA STRATEGIES:

Visualizing Data:

Visualizing data is to literally create and then consider a visual display of data. Technically, it is not analysis, nor is it a substitute for analysis. However, visualizing data can be a useful starting point prior to the analysis of data.

Involves: Creating a visual "picture" or graphic display of the data.
Reasons: a way to begin the analysis process; or as an aid to the reporting/ presentation of findings.

Exploratory Analysis:

Exploratory analysis entails looking at data when there is a low level of knowledge about a particular indicator (teacher qualifications, first and second language acquisition, etc.) It could also include the relationship between indicators and/or what is the cause of a particular indicator.

Involves: Looking at data to identify or describe "what's going on"? – creating an initial starting point for future analysis.
Reasons: Like you have a choice?

Trend Analysis:

The most general goal of trend analysis is to look at data over time. One aspect of trend analysis that is discussed and encouraged is that of comparing one time period to another time period. This form of trend analysis is carried out in order to assess the level of an indicator before and after an event.

Involves: Looking at data collected at different periods of time.
Reasons: to identify and interpret and, potentially, estimate change.

Estimation:

Estimation procedures may occur when working with either quantitative or qualitative data. The use of both quantitative data such as poverty level data, can be combined with interviews from providers serving low-income families to help determine the proportion of families in the area that are income eligible. Estimation is one of many tools used to assist planning for the future. Estimation works well for forecasting quantities that are closely related to demographic characteristics, eligible children and families, and social services. Estimation is the combination of information from different data sources to project information not available in any one source by itself.

Involves: Using actual data values to predict a future value.

Reasons: to combat boredom after you have mastered all the previous strategies. Also, to answer PIR and Community Assessment items and tasks.

CHAPTER 3
DATA COLLECTION

3.1 DATA COLLECTION

Data collection is a process in and of itself, in addition to being a part of the larger whole. Data come in many different types and can be collected from a variety of sources, including:

- Observations
- Questionnaires
- Interviews
- Documents
- Tests
- Others

In order to successfully manage the data collection process, programs need a plan that addresses the following:

• What types of data are most appropriate to answer the questions?

• How much data are necessary?

• Who will do the collection?

• When and where will the data be collected?

• How will the data be compiled and later stored?

By creating a data collection plan, programs can proceed to the next step of the overall process.

3.2 COLLECTING DATA USING API's

Data Collection is the first step in solving any analysis problem and can be collected in many formats and from many sources. In the first module of the Capstone, we will collect data by scraping the internet and using web APIs.

3.2.1 Getting Started with Watson Studio

IBM Watson Studio provides you with the environment and tools to solve problems by collaboratively working with data. You can choose the tools you need to analyze and visualize data, to cleanse and shape data, to ingest streaming data, or to create and train machine learning models. You need an IBM Cloud account to create a project in Watson Studio

3.2.2 API Access Details

How to Get Data!?

In the coming future data will increasingly be shared using Application Programming Interfaces (APIs). An API is hosted on the public internet or a private network. APIs provide anytime access to the latest data.

Your assignment is to get the number of job openings using the GitHub Jobs API for technologies like:

- Java
- MySQL
- C#
- Python
- C++

GitHub Jobs API allows anyone to query for the jobs based upon:

- Technology like Python, MySQL
- City like New York, Bangalore

Here are the technical details to access the API.

The GitHub Jobs API allows you to search, and view jobs with JSON over HTTP.

To get the JSON representation of any search result or job listing, append. json to the URL you'd use on the HTML GitHub Jobs site.

Pagination

The API also supports pagination. /Positions. Json, for example, will only return 50 positions at a time. You can paginate results by adding a page parameter to your queries. Pagination starts by Default at 0.

3.3 EXPLORING DATA

About the dataset Stack Overflow, a popular website for developers, conducted an online survey of software professionals across the world. The survey data was later open sourced by Stack Overflow. The actual data set has around 90,000 responses. The dataset you are going to use in this assignment comes from the following source: https://stackoverflow.blog/2019/04/09/the-2019stack-overflow-developer-survey-results-are-in/ under a ODbL: Open Database License. You will be given a subset of the original data set in this capstone project. You will explore, analyze, and visualize this dataset and present your analysis.

Note: This randomized subset contains around 1/10th of the original data set. Any conclusions you draw after analyzing this subset may not reflect the real-world scenario.

The dataset is available as a .csv file here.
The below table lists the questions asked in the survey and the column under which the response was collected.

Column Name	Question Text
Respondent	Randomized respondent ID number (not in order of survey response time)
Main Branch	Which of the following options best describes you today? Here, by "developer" we mean "someone who writes code."
Hobbyist	Do you code as a hobby?
Opensource	How often do you contribute to open source?
Opensource	How do you feel about the quality of open-source software (OSS)?
Employment	Which of the following best describes your current employment status?
Country	In which country do you currently reside?
Student	Are you currently enrolled in a formal, degree-granting college or university program?
Ed Level	Which of the following best describes the highest level of formal education that you've completed?
Undergrad Major	What was your main or most important field of study?
Edu Other	Which of the following types of non-degree education have you used or participated in? Please select all that apply.
Org Size	Approximately how many people are employed by the company or organization you work for?
Dev Type	Which of the following describe you? Please select all that apply.
Years Code	Including any education, how many years have you been coding?
Age1stCode	At what age did you write your first line of code or program? (E.g., webpage, Hello World, Scratch project)
YearsCodePro	How many years have you coded professionally (as a part of your work)?
Career Sat	Overall, how satisfied are you with your career thus far?

JobSat How satisfied are you with your current job? (If you work multiple jobs, answer for the one you spend the most hours on.)
MgrIdiot How confident are you that your manager knows what they're doing?
MgrMoney Do you believe that you need to be a manager to make more money?
MgrWant Do you want to become a manager yourself in the future?
JobSeek Which of the following best describes your current job-seeking status?
LastHireDate When was the last time that you took a job with a new employer?
LastInt In your most recent successful job interview (resulting in a job offer), you were asked to... (check all that apply)
FizzBuzz Have you ever been asked to solve FizzBuzz in an interview?
JobFactors Imagine that you are deciding between two job offers with the same compensation, benefits, and location. Of the following factors, which 3 are MOST important to you?
ResumeUpdate Think back to the last time you updated your resumé, CV, or an online profile on a job site. What is the PRIMARY reason that you did so?
CurrencySymbol Which currency do you use day-to-day? If your answer is complicated, please pick the one you're most comfortable estimating in.
CurrencyDesc Which currency do you use day-to-day? If your answer is complicated, please pick the one you're most comfortable estimating in.
CompTotal What is your current total compensation (salary, bonuses, and perks, before taxes and deductions), in CurrencySymbol? Please enter a whole number in the box below, without any punctuation. If you are paid hourly, please estimate an equivalent weekly, monthly, or yearly salary. If you prefer not to answer, please leave the box empty.
CompFreq Is that compensation weekly, monthly, or yearly?
ConvertedComp Salary converted to annual USD salaries using the exchange rate on 20190201, assuming 12 working months and 50 working weeks.
WorkWeekHrs On average, how many hours per week do you work?
WorkPlan How structured or planned is your work?
WorkChallenge Of these options, what are your greatest challenges to productivity as a developer? Select up to 3:
WorkRemote How often do you work remotely?
WorkLoc Where would you prefer to work?
ImpSyn For the specific work you do, and the years of experience you have, how do you rate your own level of competence?
CodeRev Do you review code as part of your work?
CodeRevHrs On average, how many hours per week do you spend on code review? UnitTests Does your company regularly employ unit tests in the development of their products?
PurchaseHow How does your company make decisions about purchasing new technology (cloud, AI, IoT, databases)?
PurchaseWhat What level of influence do you, personally, have over new technology purchases at your organization?
LanguageWorkedWith Which of the following programming, scripting, and markup languages have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the language and want to continue to do so, please check both boxes in that row.)
LanguageDesireNextYear Which of the following programming, scripting, and markup languages have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the language and want to continue to do so, please check both boxes in that row.)

DatabaseWorkedWith Which of the following database environments have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the database and want to continue to do so, please check both boxes in that row.)
DatabaseDesireNextYear Which of the following database environments have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the database and want to continue to do so, please check both boxes in that row.)
PlatformWorkedWith Which of the following platforms have you done extensive development work for over the past year? (If you both developed for the platform and want to continue to do so, please check both boxes in that row.)
PlatformDesireNextYear Which of the following platforms have you done extensive development work for over the past year? (If you both developed for the platform and want to continue to do so, please check both boxes in that row.)
WebFrameWorkedWith Which of the following web frameworks have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the framework and want to continue to do so, please check both boxes in that row.)
WebFrameDesireNextYear Which of the following web frameworks have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the framework and want to continue to do so, please check both boxes in that row.)
MiscTechWorkedWith Which of the following other frameworks, libraries, and tools have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the technology and want to continue to do so, please check both boxes in that row.)
MiscTechDesireNextYear Which of the following other frameworks, libraries, and tools have you done extensive development work in over the past year, and which do you want to work in over the next year? (If you both worked with the technology and want to continue to do so, please check both boxes in that row.)
DevEnviron Which development environment(s) do you use regularly? Please check all that apply.
OpSys What is the primary operating system in which you work?
Containers-How do you use containers (Docker, Open Container Initiative (OCI), etc.)?
BlockchainOrg-How is your organization thinking about or implementing blockchain technology?
BlockchainIs-Blockchain / cryptocurrency technology is primarily:
BetterLife-Do you think people born today will have a better life than their parents?
ITperson-Are you the "IT support person" for your family?
OffOn-Have you tried turning it off and on again?
SocialMedia-What social media site do you use the most?
Extraversion-Do you prefer online chat or IRL conversations?
ScreenName-What do you call it?
SOVisit1st-To the best of your memory, when did you first visit Stack Overflow?
SOVisitFreq-How frequently would you say you visit Stack Overflow?
SOVisitTo-I visit Stack Overflow to... (check all that apply)
SOFindAnswer On average, how many times a week do you find (and use) an answer on Stack Overflow?
SOTimeSaved-Think back to the last time you solved a coding problem using Stack Overflow, as well as the last time you solved a problem using a different resource. Which was faster?

SOHowMuchTime-About how much time did you save? If you're not sure, please use your best estimate.
SOAccount-Do you have a Stack Overflow account?
SOPartFreq-How frequently would you say you participate in Q&A on Stack Overflow? By participate we mean ask, answer, vote for, or comment on questions. SOJobs
Have you ever used or visited Stack Overflow Jobs?

CHAPTER 4
DATA WRANGLING

Data Wrangling or Munging is a process in which we clean up the data set and make it ready for data analysis. In this assignment you will perform the following tasks:

- ✦ Identify duplicate rows in the data frame.
- ✦ Remove duplicate rows from the data frame.
- ✦ Find the number of missing values for all columns.
- ✦ Find the value counts for the column "Employment".
- ✦ Normalize the data using two existing columns.

4.1.1 STEPS PERFORMED:

- **Identify duplicate rows in the data frame:**

Syntax: DataFrame.duplicated(subset = None, keep = 'first') **Parameters:**

subset: This Takes a column or list of column label. Its default value is None. After passing columns, it will consider them only for duplicates.

keep: This Controls how to consider duplicate value. It has only three distinct value and default is 'first'.

- If 'first', This considers first value as unique and rest of the same values as duplicate.
- If 'last', This considers last value as unique and rest of the same values as duplicate.
- If 'False', This considers all of the same values as duplicates. **Returns:**

Boolean Series denoting duplicate rows.

- **Removing duplicate rows from the dataframe:**

Pandas is one of those packages and makes importing and analyzing data much easier. An important part of Data analysis is analyzing Duplicate Values and removing them. Pandas drop_duplicates() method helps in removing duplicates from the data frame.

Syntax: DataFrame.drop_duplicates(subset=None, keep='first', inplace=False) **Parameters:**

subset: Subset takes a column or list of column label. Its default value is none. After passing columns, it will consider them only for duplicates. **keep**: keep is to control how to consider duplicate value. It has only three distinct value and default is 'first'.

- If 'first', it considers first value as unique and rest of the same values as duplicate.
- If 'last', it considers last value as unique and rest of the same values as duplicate.
- If 'False', it considers all of the same values as duplicates **inplace:** Boolean values, removes rows with duplicates if True.

Return type: DataFrame with removed duplicate rows depending on Arguments passed.

- **Find the number of missing values for all columns:**

Dealing with missing values is one of the common tasks in doing data analysis with real data. A quick understanding on the number of missing values will help in deciding the next step of the analysis.

We will use Pandas's isna() function to find if an element in Pandas dataframe is missing value or not and then use the results to get counts of missing values in the dataframe. When applied to a dataframe, Pandas isna() function return Boolean dataframe with True with the element is missing value and False when it is not a missing value.

We can use Pandas' sum () function to get the counts of missing values per each column in the dataframe. By default, Panda's sum () adds across columns. And we get a dataframe with number of missing values for each column.

- **Find the value counts for the column "Employment":**

counting number of Values in a Row or Columns is important to know the Frequency or Occurrence of data. First find out the shape of dataframe i.e., number of rows and columns in this dataframe.

df.shape

Here's how to count occurrences (unique values) in a column in Pandas dataframe:

Pandas count distinct values in column df['employment'].value_counts()

Running the value_counts method on the DataFrame (rather than on a specific column) will return the number of unique values in all the DataFrame columns.

- **Normalize the data using two existing columns:**

Data Normalization:

Data Normalization could also be a typical practice in machine learning which consists of transforming numeric columns to a standard scale. In machine learning, some feature values differ from others multiple times. The features with higher values will dominate the learning process.

Using The maximum absolute scaling:

The maximum absolute scaling rescales each feature between -1 and 1 by dividing every observation by its maximum absolute value. We can apply the maximum absolute scaling in Pandas using the .max () and. abs () methods.

Using The min-max feature scaling:

The min-max approach (often called normalization) rescales the feature to a hard and fast range of [0,1] by subtracting the minimum value of the feature then dividing by the range.
We can apply the min-max scaling in Pandas using the. min () and .max () methods.

Using The z-score method:

The z-score method (often called standardization) transforms the info into distribution with a mean of 0 and a typical deviation of 1. Each standardized value is computed by subtracting the mean of the corresponding feature then dividing by the quality deviation.

Using sklearn:

Transform features by scaling each feature to a given range. This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g., between zero and one.

CHAPTER 5
Exploratory Data Analysis

After cleaning the dataset, next step is the analysis. In this stage we will become more familiar with the data set and it will start to take shape.

Exploratory Data Analysis is a method of analyzing data sets by summarizing their main characteristics with visualizations. It not only uncovers the hidden trends and insights but also leads us towards building an accurate model for prediction.

5.1 Understanding data:

Quality data is fundamental and plays a significant role in the prediction. In order to get the best results, the appropriate data must be sourced and understood.

Variable identification and understanding:

• Understand the variables and the type of data for each variable.

• Identify predictor variables, target variables, the data type of variables

Univariate Analysis

• Univariate analysis is the simplest form of analyzing data.

• "Uni" means "one"; therefore, we analyze one variable at a time.

• It is mainly used to describe, summarize and find patterns in the data.

• It is also helpful to deal with missing values and handle outliers.

Visualization techniques used:

a) count plot:

A count plot is a histogram plotted across a category variable rather than a quantitative variable.

b) Dist plot:

A histogram with a line can be displayed using Seaborn distplot. Seaborn is used in conjunction with matplotlib, a Python plotting library. A distplot is a graph that shows a singlevariate distribution of observations. The matplotlib hist function is combined with the seaborn kdeplot() and rugplot() functions in the distplot() function.

1. Plot a distribution curve, and histogram.
2. Find the median, and outliers of particular columns.
3. Compute the Inter Quartile Range.
4. Find out the upper and lower bounds, and find correlations between numerical columns.
5. Create a new dataframe.

5.2 Identifying Outliers in the dataset:

Outlier detection and handling:

With some models, outliers can cause issues. Linear regression methods, for example, are less resistant to outliers than decision tree models. In general, we should not delete outliers unless there is a compelling reason to do so. Removing them increases performance in certain cases but not in others. As a result, an outlier must be removed for a valid cause, such as suspicious observations that are unlikely to be part of true data

- Sorting Your Datasheet to Find Outliers:

Sorting your datasheet is a simple but effective way to highlight unusual values. Simply sort your data sheet for each variable and then look for unusually high or low values.

- Graphing Your Data to Identify Outliers

Boxplots, histograms, and scatterplots can highlight outliers.
Boxplots display asterisks or other symbols on the graph to indicate explicitly when datasets contain outliers. These graphs use the interquartile method with fences to find outliers, which I explain later. The boxplot below displays our example dataset. It's clear that the outlier is quite different than the typical data value.

- Using Z-scores to Detect Outliers

Z-scores can quantify the unusualness of an observation when your data follow the normal distribution. Z-scores are the number of standard deviations above and below the mean that each value falls. For example, a Z-score of 2 indicates that an observation is two standard deviations above the average while a Z-score of -2 signifies it is two standard deviations below the mean. A Z-score of zero represents a value that equals the mean.

To calculate the Z-score for an observation, take the raw measurement, subtract the mean, and divide by the standard deviation. Mathematically, the formula for that process is the following:

$$Z = \frac{X - \mu}{\sigma}$$

The further away an observation's Z-score is from zero, the more unusual it is. A standard cut-off value for finding outliers are Z-scores of +/-3 or further from zero. The probability distribution below displays the distribution of Z-scores in a standard normal distribution. Z-scores beyond +/- 3 are so extreme you can barely see the shading under the curve.

- Using the Interquartile Range to Create Outlier Fences

You can use the interquartile range (IQR), several quartile values, and an adjustment factor to calculate boundaries for what constitutes minor and major outliers. Minor and major denote the unusualness of the outlier relative to the overall distribution of values. Major outliers are more extreme. Analysts also refer to these categorizations as mild and extreme outliers.

The IQR is the middle 50% of the dataset. It's the range of values between the third quartile and the first quartile (Q3 – Q1). We can take the IQR, Q1, and Q3 values to calculate the following outlier fences for our dataset: lower outer, lower inner, upper inner, and upper outer.

These fences determine whether data points are outliers and whether they are mild or extreme.

Values that fall inside the two inner fences are not outliers

5.2.1 Remove Outliers from data:

An outlier of a dataset is defined as a value that is more than 3 standard deviations from the mean. Removing outliers from a pandas.DataFrame removes any rows in the DataFrame which contain an outlier. Outlier calculations are performed separately for each column.

USE scipy.stats.zscore() to remove outliers from a dataframe

Call scipy.stats.zscore(a) with a as a DataFrame to get a NumPy array containing the z-score of each value in a. Call numpy.abs(x) with x as the previous result to convert each element in x to its absolute value. Use the syntax (array < 3). all(axis=1) with array as the previous result to create a boolean array. Filter the original DataFrame with this result.

5.3 Identify correlation between features in the dataset:

What is correlation?

Correlation is a statistical term which in common usage refers to how close two variables are to having a linear relationship with each other.

Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, we can drop one of the two features

In pandas you can easily get the correlation matrix by using: corrMatrix = df.corr()

CHAPTER 6
DATA VISUALIZATION

Introduction:

6.1 What is Data Visualization?

Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals. The resulting visual representation of data makes it easier to identify and share real-time trends, outliers, and new insights about the information represented in the data.

A dashboard is an information visualization tool. It helps you monitor events or activities at a glance by providing insights on one or more pages or screens. Unlike an infographic, which presents a static graphical representation, a dashboard conveys real-time information by pulling complex data points directly from large data sets. An interactive dashboard makes it easy to sort, filter, or drill into different types of data as needed. Data science techniques can be used to identify what is happening, why it's happening, and what will happen next at speed.

As the amount of big data increases, more people are using data visualization tools to access insights on their computer and on mobile devices. Dashboards are used by business people, data analysts, and data scientists to make data-driven business decisions.

6.1.1 WHY IT IS IMPORTANT ...!

- See the big picture
- Make better decisions
- Present meaningful data
- Democratize your data

- Uncover insights and see patterns within complex data without relying on a data scientist.
- Understand your next steps and spend less time performing data analysis. Quickly act on decisions.
- Share insights with others in an easy-to-understand form –

Provide one source of truth for your entire organization.

6.2 CHARTS:

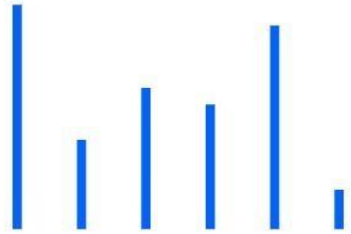
Charts are typically divided into categories based on their goals, aesthetics or visual features. Since charts can be versatile and used in different ways, details and features of these categories are explained and contextualized here.

- ✦ Comparisons
- ✦ Trends
- ✦ Part to whole
- ✦ Correlations
- ✦ Relationships and connections
- ✦ Maps

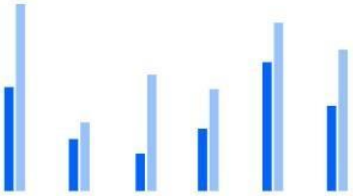
Comparisons:

Charts designed for comparison aim to visualize differences between elements. Most of the time comparisons rely on the ability of the human eye to identify longer or bigger shapes with very little or no effort. Side-by-side positioning and alignment of the visual elements make comparisons even easier. These charts are used for time-based data, for example, units sold per day or worked hours per month. They are also used for categorized data, for example, revenue by market or sold units by team.

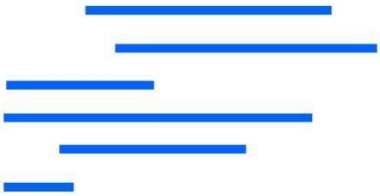
Simple bar



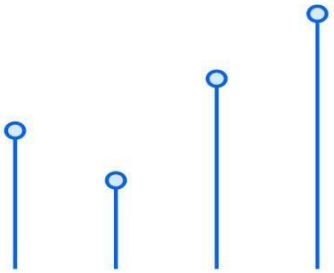
Grouped bar



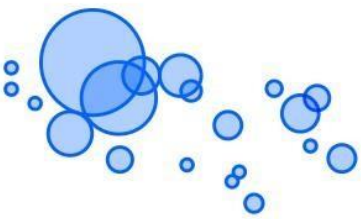
Floating bar



Lollipop



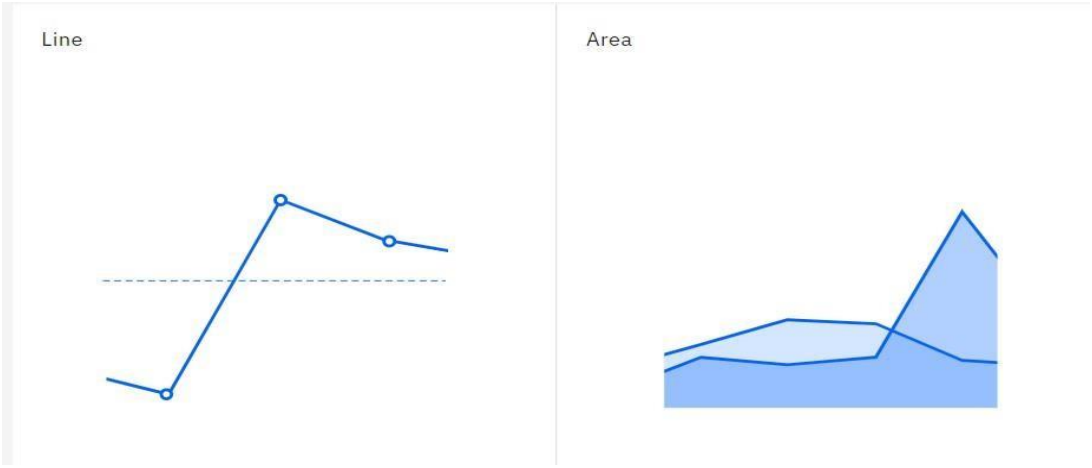
Bubble

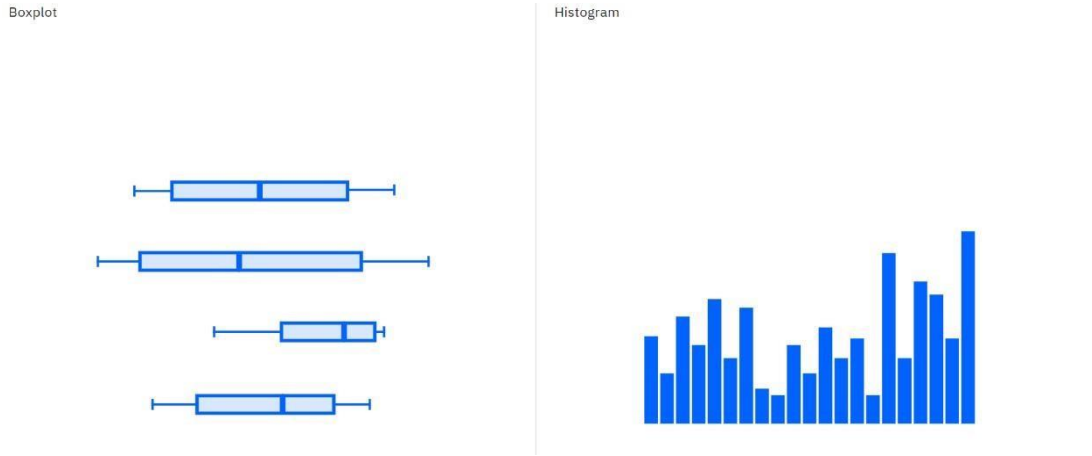


Wordcloud

Trends:

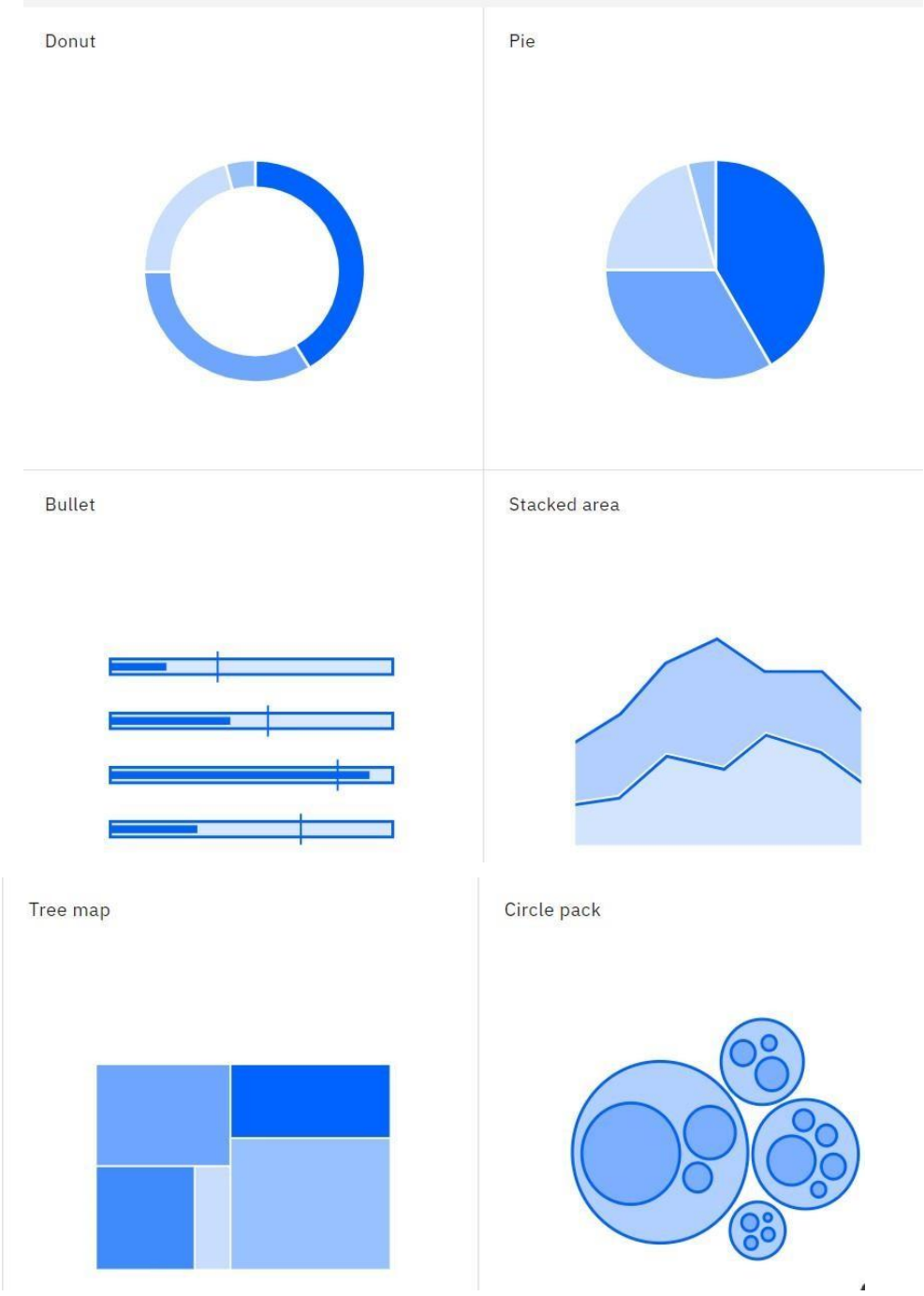
Trend charts represent data along with the time dimension. Use them mainly to track changes over periods of time of varying duration and scale. They rely on direction to show the evolution of consecutive values and might be influenced by different cultural contexts. These charts are used for time-based data, for example, revenue by quarter or rainfall per day.





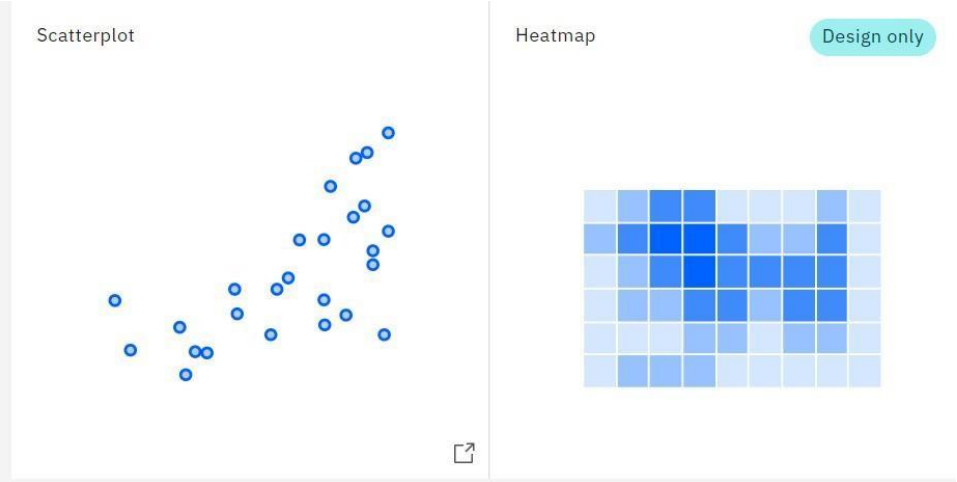
Part to whole:

The goal of these charts is to show the inner subdivision of a value among different categories or groups. Mostly used to represent percentages, they can also be used for absolute values. Their function does not depend on the graphic shapes used, such as pie, donut, square and so on. These charts are used for categorized data, for example, subdivision of revenue by product or percentage of users by browser.



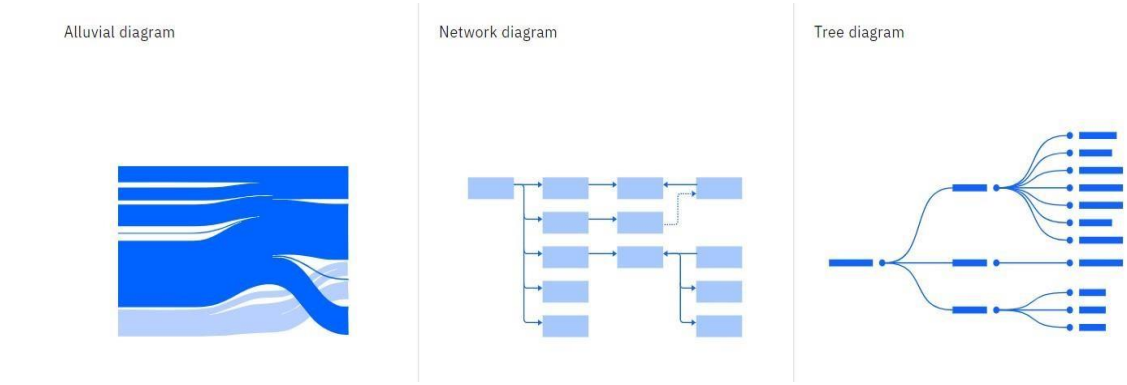
Correlations:

These charts are better suited to highlight the possible correlation between two or more indicators and how they might affect each other. Correlation charts have the final goal of making it easier for the human eye to spot combined behaviors. These charts are used for multidimensional data, for example, correlation between phone-call duration and customer satisfaction.



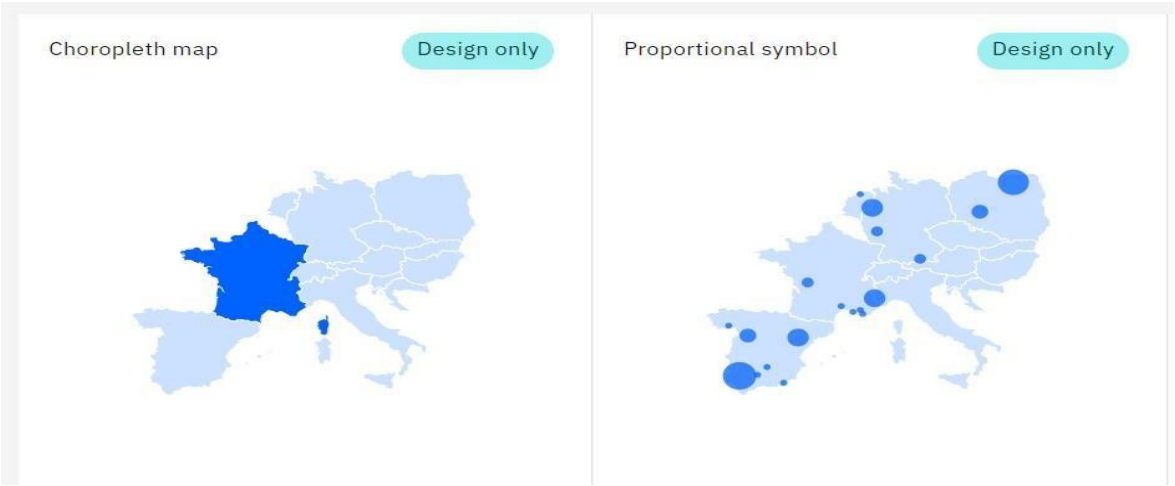
Relationships and connections:

Charts included in this category represent hierarchies. The intent is to explain the role of an element within an ecosystem or to observe the inner nature of a subject in different phases and states of a process. These charts are used for categorized data, for example, country of origin of asylum seeker and gender. They are also used for multidimensional data, for example, number of active users by testing phase.



Maps:

Maps are the easiest and most immediate way to communicate geolocated information. Maps allow the user to recognize areas and places, to understand the geographical context of the topic and to identify patterns, all relying on the position of elements. These charts are used for geographical data, for example, voters by county or average wage by neighborhood.



6.3 TO CREATE VISUALIZATION:

- Define your intent with users
  - Identify who can benefit from data visualization. Rely on simple charts up front.
- Understand and clean data
  - Look at your data set structure typologies. Analyze rows and columns for inconsistencies.
- Model data, check for visual validity
  - Use basic visual models to see and understand enormous data sets. ID patterns and trends.
- Experiment with structure and style
  - Draft different design versions. Try a variety of charts while staying within your established organizational graphic standards.
- Test and iterate
  - Gather user impressions and opinions. Use research to influence visualization method iterations and justify changes.
- Refine and implement
  - Look for bugs and functional errors. Check your visualizations for inconsistencies.

6.3.1 Benefits of good data visualization:

The basic uses of the Data Visualization technique are as follows:

- It is a powerful technique to explore the data with presentable and interpretable results. - In the data mining process, it acts as a primary step in the pre-processing portion.
- It supports the data cleaning process by finding incorrect data and corrupted or missing values.
- It also helps to construct and select variables, which means we have to determine which variable to include and discard in the analysis.
- In the process of Data Reduction, it also plays a crucial role while combining the categories.

CHAPTER 7
Dashboard

In this we used IBM Cognos Dashboard Embedded (CDE) is an AI-fueled business intelligence service that supports the entire data analytics cycle, from discovery to operationalization. It provides users with data discovery capabilities to visually explore and interact with their data to identify the key insights for improving data driven decisions. Users can perform data discovery and then quickly assemble that information into interactive, visually appealing dashboards; all without the need of formal training.

Dataset Used in this Lab:

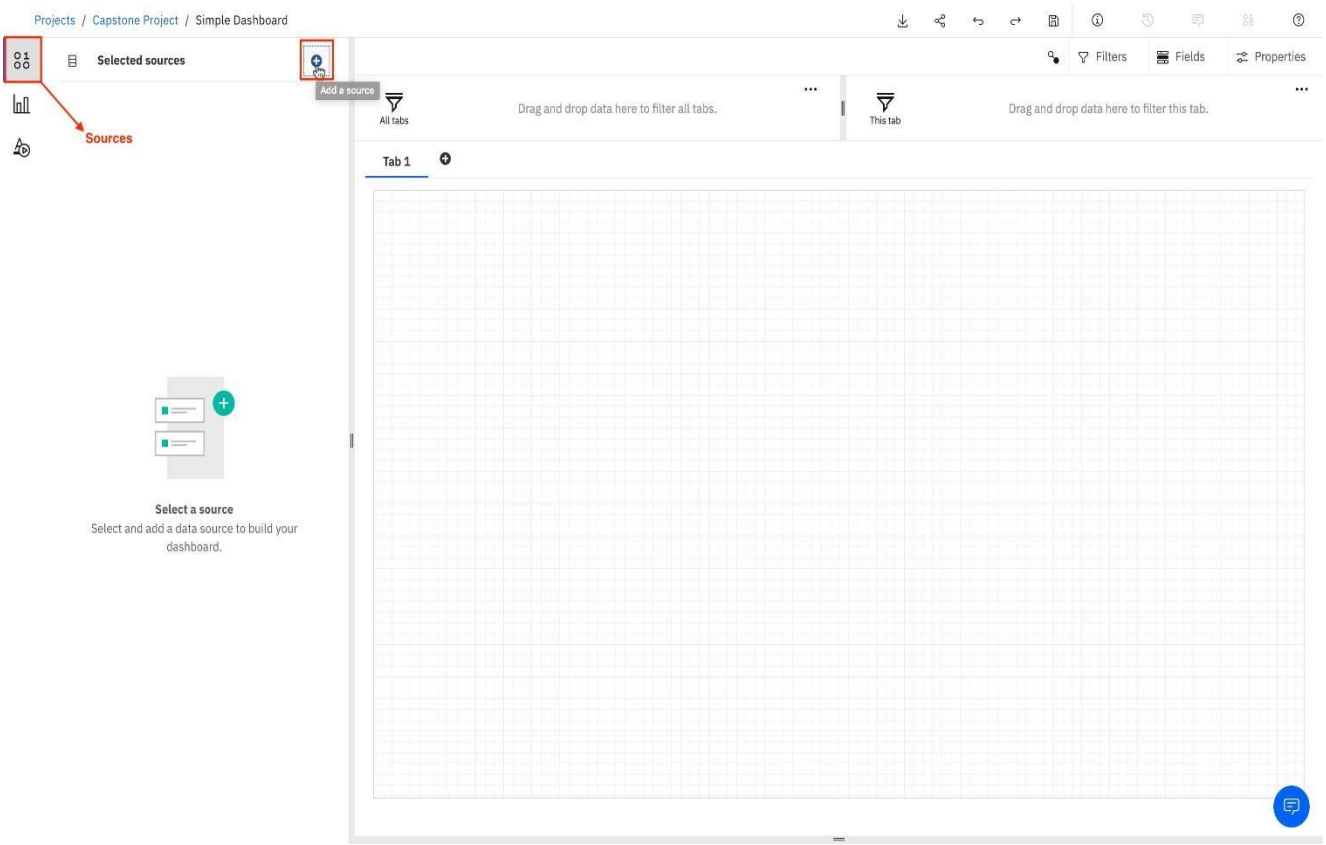
The dataset used in this lab comes from the following source:

https://www.kaggle.com/kyanyoga/sample-sales-data under a **CC0: Public Domain license**.

Objectives:

- Login to IBM Cloud Pak for Data platform through IBM Cloud
- Create a project in IBM Cloud Pak for Data
- Add a Cognos Dashboard Embedded (CDE) service to your created project
- Navigate around the Cognos Dashboard Embedded (CDE) user interface
- Upload external data files to your created project (Supports .CSV files only)
- Start a new dashboard with a dashboard template and populate it with a data visualization.

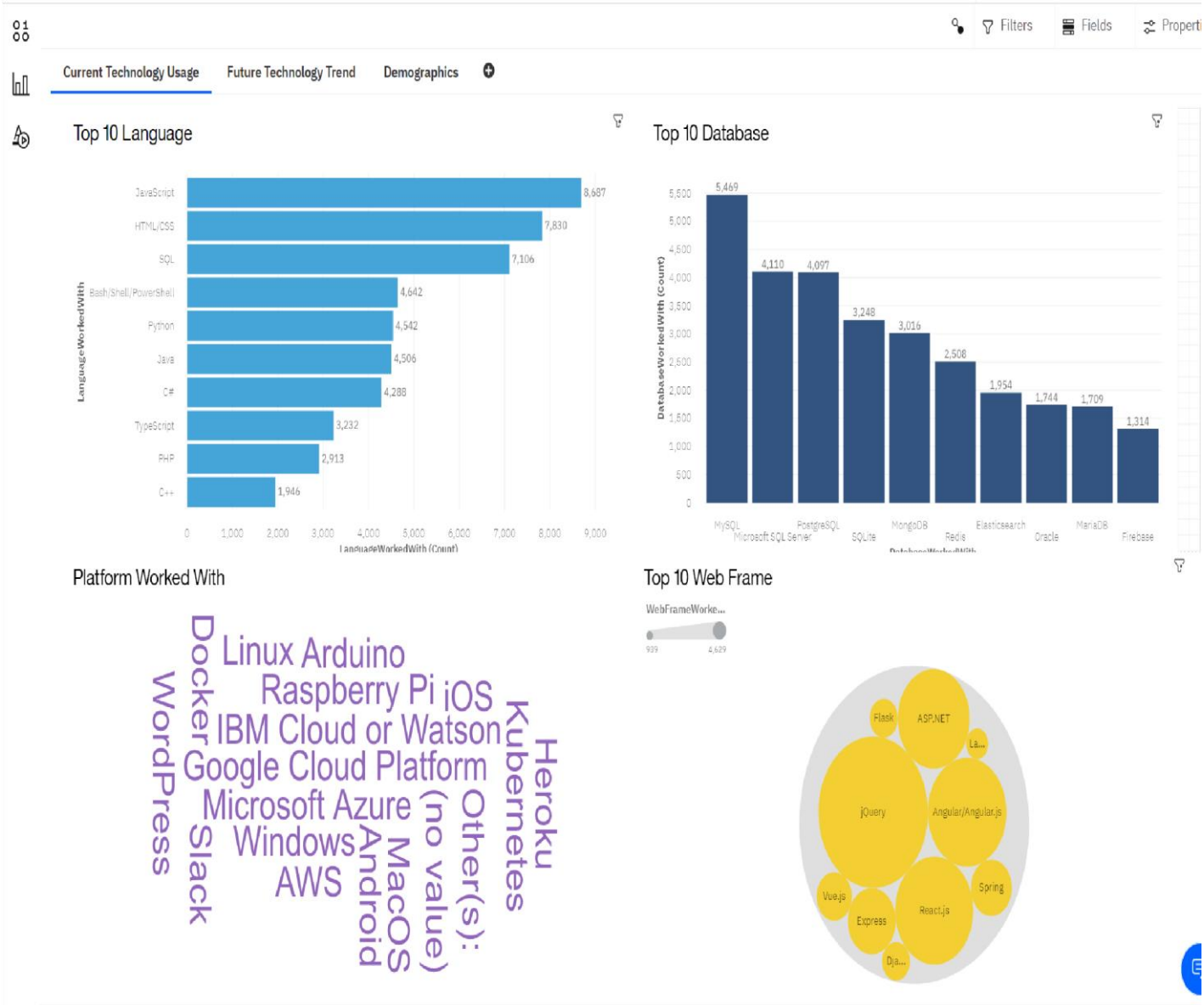
→ Sample Dashboard after creating is shown as below



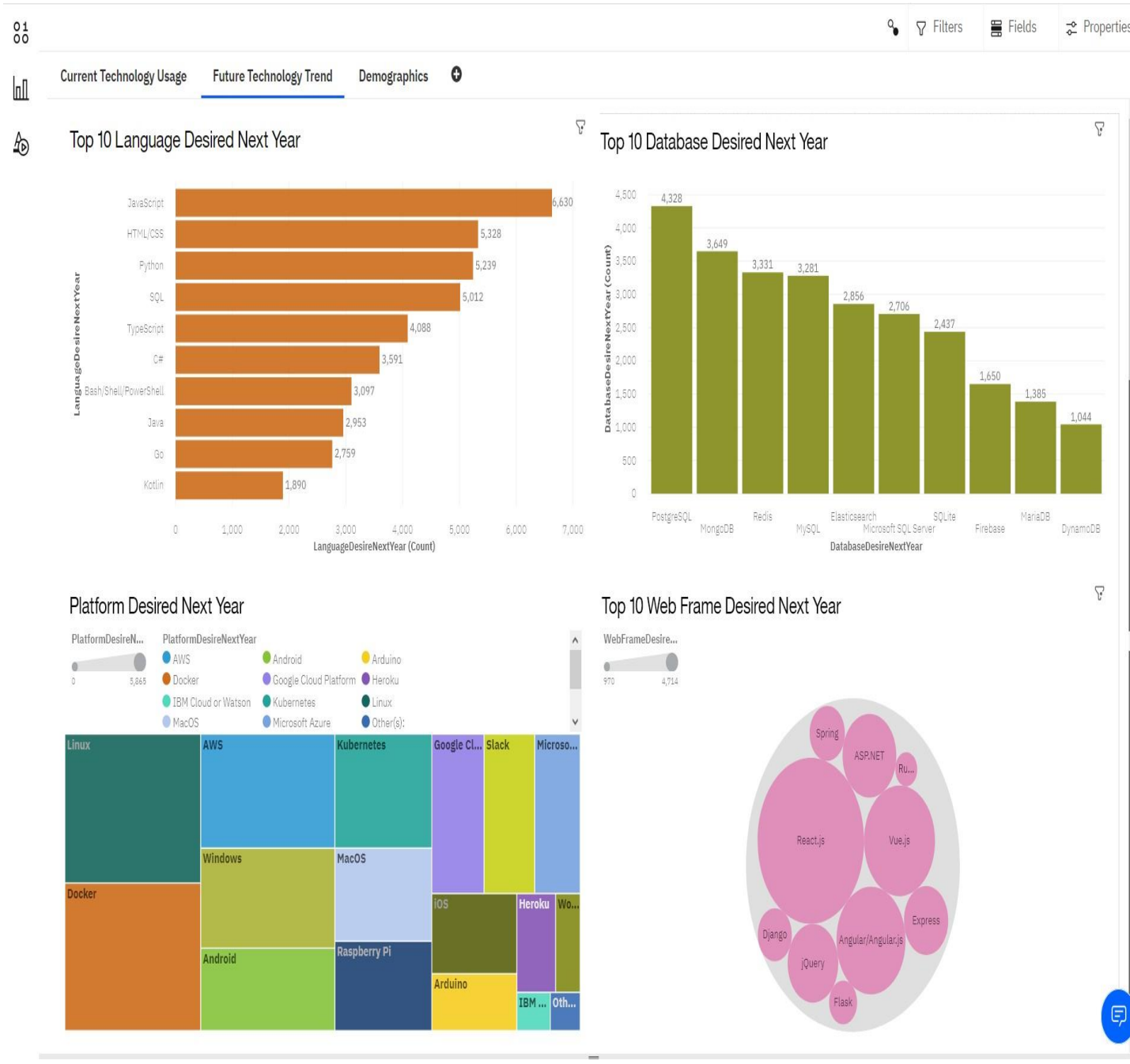
→ Create 3 dashboards (3 separate tabs under a single dashboard) as follows:

- One dashboard using the 2 x 2 rectangle areas tabbed template - rename this dashboard tab to **Current Technology Usage**.
- One dashboard using the 2 x 2 rectangle areas tabbed template - rename this dashboard tab to **Future Technology Trend**.
- One dashboard using the 2 x 2 rectangle areas tabbed template - rename this dashboard tab to **Demographics**.
-

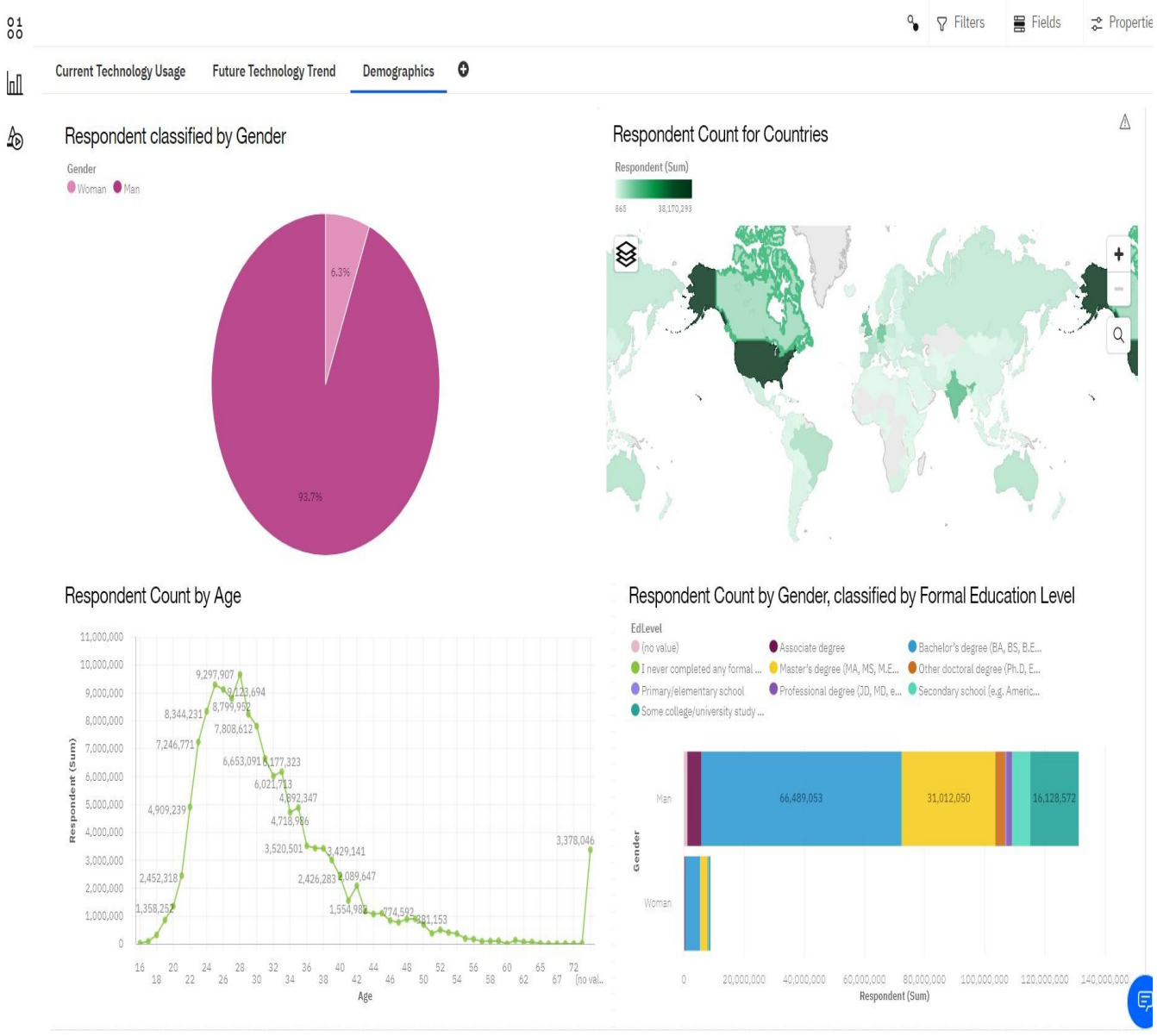
- **Visualizations of current technology usage based on the required metrics.**



- **Visualization of Future technology trend based on required metrics.**



- **Visualization of demographics based on required metrics**



CHAPTER 8
CONCLUSION

8.1 Overall Findings and Implications:

Findings:

- Fast changing technology every year
- Concentration on several countries like USA and India
- Gender gap in technology jobs
- Platforms like Docker and AWS are growing

Implications:

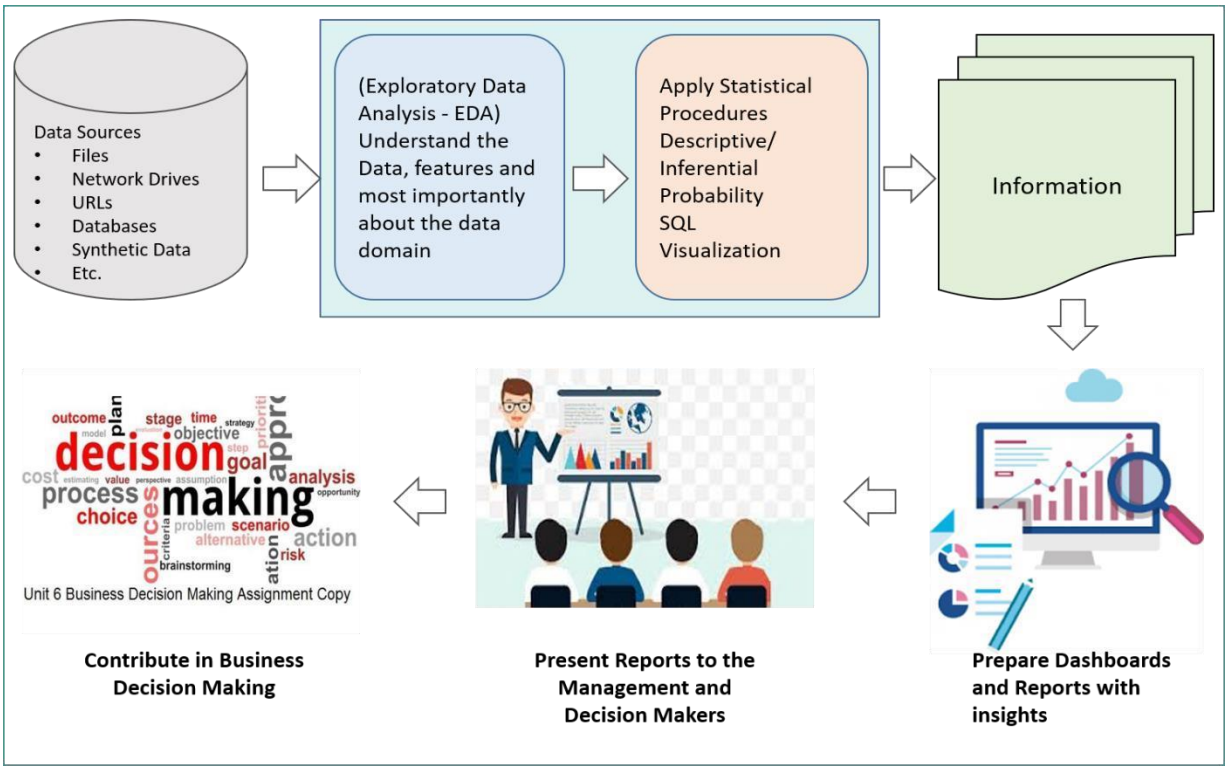
- Companies need to be flexible and adjust to rapid changes
- Need to spread technology out to lagging countries
- Impact of job hiring's
- Shift to faster app deployments and cloud services in future

CHAPTER 9
ROLES AND RESPONSIBILITIES

9.1 Data analyst role:

Data analysts mostly work with an organization's structured data. They create reports, dashboards, and other visualizations on data associated with customers, business processes, market economics, and more to provide insights to senior management and business leaders in support of decision-making efforts. Data analysts work with all manner of data, including inventories, logistics and transportation costs, market research, profit margins, sales figures, and so on. They use this data to help the business estimate market share, price products, time sales, optimize transportation costs, and the like.

- Determine Organizational Goals by understanding the Business requirements from IT and Business needs
- Data Analysts have to often mine or collect data. Getting data from the company database or extracting it from external sources to do any sort of research is one of the major roles of any Data Analyst.
- Data Analysts must start with a thorough data cleansing process. The good analysis rests on clean data–it’s as simple as that. Cleaning involves removing data that may distort your analysis or standardizing your data into a single format.
- While analyzing the data, they follow the process of evaluating data using analytical and logical reasoning to examine each component of the data provided. There are various tools and programming languages used in the analysis.
- Data analyst spend a significant part of time on finding trends, correlations, and patterns in the complicated datasets. Trends are also important. Data Analysts look for both short-term and long-term trends. It helps you understand how your business has performed and predict where current business operations and practices will take you. It will give you ideas about how you might change things to move your business in the right direction.
- Being able to tell a compelling story with data is crucial to getting your point across and keeping your audience engaged. For this reason, data visualization can have a makeorbreak effect when it comes to the impact of your data. Analysts use eye-catching, highquality charts and graphs to present their findings in a clear and concise way.


Data Sources
Files
Network Drives
URLs
Databases
Synthetic Data
Etc.
(Exploratory Data Analysis - EDA) Understand the Data, features and most importantly about the data domain
Apply Statistical Procedures
Descriptive/ Inferential
Probability
SQL
Visualization
Information
Prepare Dashboards and Reports with insights
Present Reports to the Management and Decision Makers
Contribute in Business Decision Making
Unit 6 Business Decision Making Assignment Copy

9.2 Data analyst responsibilities:

Data analysts seek to understand the questions the business needs to answer and determine whether those questions can be answered by data. They must understand the technical issues associated with collecting data, analyzing data, and reporting. They must be able to recognize trends and patterns. According to Workable, key data analyst responsibilities include:

- Analyzing data using statistical techniques and providing reports
- Developing and implementing databases and data collection systems
- Acquiring data from primary and secondary sources and maintain data systems
- Identifying, analyzing, and interpreting trends or patterns in complex data sets
- Filtering and cleaning data
- Working with management to prioritize business and information needs
- Locating and defining new process improvement opportunities
- Extract actionable insights from large databases
- Perform recurring and ad-hoc quantitative analysis to support day-to-day decision making
- Create data dashboards, graphs and visualizations
- Prepare reports for internal and external audiences using business analytics reporting tools
- Help translate data into visualizations, metrics, and goals
- Write SQL queries to extract data from the data warehouse
- Identify areas to increase efficiency and automation of processes
- Set up and maintain automated data processes
- Identify, evaluate and implement external services and tools to support data validation and cleansing
- Gather, understand and document detailed business requirements using appropriate tools and techniques