**Store Sales**

**This Case Study has three check points defined in it.**

| **Check Point Topics** | **Remarks** | **Max Marks** |
| --- | --- | --- |
| 1.1 Data manipulation using Python ( 25 marks)  1.2 Analysis using SQL Queries (25 Marks)  1.3 Statistical Analysis using Python (25 Marks) | **Check point 1** | **75** |
| 2.1 Visualization using Python(20 marks)  2.2 Exploratory Data Analysis(40 marks)  2.3Visualization using Power-BI (25 marks)  2.4 - Model Building using ML algorithms (40 marks) | **Check Point 2** | **125** |
| 3.1 Data Analysis using Big Data Tools(35 marks)  3.2 Data Analysis on Cloud (35 marks)  3.3 Deployment of ML model using Flask (30 marks) | **Check point 3** | **100** |

**Domain:**

**Retail and Supply Chain**

**About:**

Dean’s is a large web e-tailer chain which sells different types of products nationally. It has shipping services via different mode/types of transportation. It keeps track of order details, shipping details, discounts provided to the customer, profit earned etc.

**Challenges:**

Currently, category managers for different categories manage budgets and marketing campaigns, and there is no cohesive strategy. Management wants to assess what is the current state of the portfolio, in terms of product mix, demand trends across categories and products and understand if there are any regional variations, and other patterns that may help in creating a marketing strategy that is both consistent at the company level but can also support variations by category, product, region, and customer type.

The first step of this extended analysis therefore is to assess what data is available and perform some exploratory and descriptive analytics to identify interesting and useful patterns, trends, and insights.

**What is Expected?**

Being a data analyst, you must come up a first step document that lists output of your exploratory analysis, any issues or problems you may see with data that need follow up, and some basic descriptive analysis that you think highlights important outcomes/findings from the data. Based on your findings, the next level of analysis will be charted out.

**Data Dictionary:**

* RowID: Row Number indicating transactions
* OrderID: A code indicating order number
* Order Date: Date on which a product has been ordered by the customer
* Ship Date: Date on which the product is shipped to the customer
* Ship Mode: The mode of shipment
* Customer ID: A unique ID given to the customer
* Customer Name: Name of the customer
* Segment: The type of segment to which the customer belongs to
* Country: Name of the country to which the product to be shipped
* City\_Code : Code of the city to which the product to be shipped
* State\_Code: Code of the State to which the product to be shipped
* Postal\_Code: Postal code of the area to which the product to be shipped
* Region\_ID: Code of the region to which the customer belongs
* Product ID: ID of the product being ordered
* Category: The category of the product
* Sub-Category: The sub-category of the product under main category
* Product Name: Name of the product being ordered
* Sales: The amount specifying the sales value
* Quantity: The quantify of the product being ordered
* Discount: The value of the discount for a specific product, if any
* Profit: The amount of profit made by the said sales

**Check Point 1**

**Task 1.1 (Data Manipulation using Python)**

Here are some indicative types of analysis you can perform. Please note that this is not an exhaustive list, you may add more

1. Check for missing values in the dataset.
2. Find out the outliers in numerical columns.
3. Product Category with highest and lowest sales.
4. The category of the product which is being sold with highest discounts
5. Percentage share of sales at every state and region.

**Task 1.2 (SQL-Oracle)**

Stage 1:

1. Construct and ER-Diagram for the above-mentioned Requirement

2. Construct Tables as per the ER-Diagram.

3. Identify the relationships between tables and use appropriate standards for the same where applicable

4. Insert the appropriate data into the identified tables from the sample dataset provided.

Stage 2:

1. Display The category name of the products that earns highest profit

2. Display Most valuable customer name who has frequently ordered products

3. Display The segment and product name which yields highest profit

4. Display name of the City with highest sales

5. Display Percentage share of profit of every region

6. Display Percentage share of sales at every state

7. Display the Category of the product with highest demand in each region

**Task 1.3 (Statistical Analysis using Python)**

* Descriptive statistics for both numerical and categorical and draw few insights from them.
* Perform relevant hypothesis testing (t, chi-Square, Anova tests)

**Check point 2 (Visualization on PowerBI, SQL, Machine Learning)**

**TASK 2.1 (Visualization using Python)**

Here are some indicative types of visualization you can perform. Please note that this is not an exhaustive list, you may add more

Come up with appropriate results and visuals for the following:

* Product Analysis:
  + Different types of product categories
  + Various sub-categories of products
  + Product with highest demand in each segment
  + Category of the product with highest demand in each region
  + The category of the product which is being sold with highest discounts
* Sales/Profit Analysis
  + The category of the product that earns highest profit
  + Most valuable customer
  + The segment which yields highest profit
  + City with highest sales
  + Percentage share of profit of every region
  + Percentage share of sales at every state

**NOTE:** Results and graphs must be backed with appropriate inferences and insights.

**TASK 2.2 (Exploratory Data Analysis)**

Data Preparation/Analysis tasks including (but not limited to) the following.

* Univariate, Bi- Variate Analysis and Multi- Variate Analysis
* Missing values identification and treatment
* Outlier analysis and treatment
* Data scaling using min-max and/or  Z-score normalisation
* Data transformation
* Feature Engineering

**TASK 2.3(Visualization using Power-BI)**

**Connect the data with Power BI desktop and perform Data Manipulation using Power Query Editor. Perform the below tasks in Power BI Desktop.**

Come up with appropriate results and visuals for the following: Please note that this is not an exhaustive list, you may add more

* Determine which State in the Central Region has the highest sales.
* Identify the City with Highest Sales in California.
* In which Region do all Product Categories fall beneath the overall average profit?
* Find the top 10 Product Names by Sales within each region
* Product with highest demand in each segment
* Which product is ranked #2 by Sales in the West region?
* Trend in profit/sales over region.
* Product with highest demand in each segment.

**NOTE:** Results and graphs must be backed with appropriate inferences and insights.

**Task 2.4(Model building using ML algorithms)**

You have to build a time series model by aggregating daily data into weekly or monthly data.

Apply various time series models like Moving Average, Exponential smoothing,

* Compare various Time series models
* Evaluate the performance of the model.
* Identify the right metric to evaluate the performance of the model.
* Identify issues and concerns on the given data and suggest the best technique/s to overcome the issues.

**Recommendations:**

* As a data analyst, what are the approaches do you suggest the sales team to forecast the sales more accurately? Recommend based on your analysis.

**NOTE:** Results and graphs must be backed with appropriate inferences and insights.

**CheckPoint 3**

**Task 3.1**

**What is Expected?**

Big Data technologies like HDFS, Hive and PySpark need to be used as the historical data increases in size. As part of this task the following activities need to be done.

* Develop a PySpark application to load data Spark DataFrames and save it into Hive tables on a Hadoop cluster in Parquet format.
* Perform profiling of the data through PySpark and ensure that it is migrated correctly wherever the source is an RDBMS.
* Write PySpark routines to cleanse the data, prepare the data to handle missing values, and the data transformations identified in task 1.1 again making sure that the data is written into Hive tables in an efficient format.
* If the predictive model identified in task 2.4 available in Spark MLlib then develop a PySpark application to implement and evaluate the ML model with appropriate metrics.
* Ensure that the best practices are followed and the design & code use the features of Spark and take advantage thereof.

**Task 3.2**

**AWS**

1. Move the Datasets to AWS s3
2. Create Redshift Instance
3. Ensure you create required tables in Redshift
4. Create a data pipeline/copy command to move the data from storage to data warehouse(Redshift). You are allowed to use other copy commands as well to move the data from storage to data warehouse.
5. Connect the Redshift data to PowerBI
6. Perform the tasks mentioned in Task 2.3(Only 4 core reports)

**AZURE**

1. Move the Dataset to Azure Synapse Storage Gen1
2. Create a serverless SQL pool to query the data from Storage gen1
3. Create a Linked service to PowerBI
4. Ensure your have sufficient privileges on Synapse to access the serverless sql pool.
5. Perform various analytics on PowerBI
6. Create the dashboard in PowerBI like Task 2.3(Only 4 core reports)

**Task 3.3**

Deploy the Machine Learning Model created Task 2.4 in Flask application