**Part 3:**

**Question 7:**

**Problem Statement:** Perform a statistical analysis on a given dataset to identify significant

trends and correlations. Provide a summary of your findings.

**Dataset:**

Employeedata.csv file is the dataset for the trends and correlation analysis.

It is also used for the statistical analysis and the hypothesis testing.

**Approach:**

**Data Understanding:** Identify the types of variables (continuous, categorical) and their distributions.

**Data Pre-processing:**

* Handle missing values with mean/median/mode.
* Treat Outliers using IQR or Isolation Forest from sklearn library.
* Identify Skewness in the dataset and treat skewness with appropriate data transformations, such as log transformation (which is best suited to transform target variable-train, predict and then reverse transform it back to original scale)
* Encode categorical variables using label encoding

**Hypothesis:**

* we can do hypothesis testing for statistical analysis for columns
* Two types of hypothesis . they are NULL hypothesis and alternate hypothesis

**Hypothesis testing:**

* it can be done by ttest,test anova or chi2
* Finding:
* Columns ExperienceInCurrentDomain and Sal are done with ttestt and resulted with alternate hypothesis

**correlations:**

correlations, only numerical data are involved. So transforming using label encoding can be done.

**Heatmap:**

Heatmap is visualized for the features correlations.

**Solution:**

Here in the taken dataset Age and Salary are highly correlated and so either one column can be used for building the model.

**Code Link:**

**Question 8:**

**Problem Statement:** Build a predictive model to forecast sales for the next quarter using

historical sales data. Explain the steps taken and the rationale behind your model choice.

**Dataset:**

Sales.csv is the dataset used for the prediction. The prediction of sales in the year 2018 is done by training the algorithm models with previous years’ data (2015,2016,2017)

**Approach:**

1. Data Understanding: Identify the types of variables (continuous, categorical) and their distributions.
2. Data Pre-processing:

* Handle missing values with mean/median/mode.
* Treat Outliers using IQR or Isolation Forest from sklearn library.
* Identify Skewness in the dataset and treat skewness with appropriate data transformations, such as log transformation (which is best suited to transform target variable-train, predict and then reverse transform it back to original)

1. Encode categorical variables using label encoding
2. EDA: Try visualizing outliers and skewness (before and after treating skewness) using Seaborn’s boxplot
3. Feature Engineering: Heatmap is visualized for the correlations of the columns
4. Model Building and Evaluation:

* Split the dataset into training and testing set
* Sales of year 2018 is for forecasting the prediction
* Sales of 2015 ,2016, 2015 are used for training

|  |  |
| --- | --- |
| **Algorithm** | **Mean Absolute Percentage error** |
| Linear Regression | 12.31 |
| Random Forest Regressor | 3.13 |
|  |  |

**Solution:**

Prediction of the future sales is done using minimal error value using the Random Forest Regressor algorithm.

**Code Link:**

**Question 9**:

Imagine you are given a large dataset with customer transactions. How would you

approach the task of identifying key customer segments and their behaviors? Describe the

steps and tools you would use.

**Approach:**

To identify key customer segments and their behaviors from a large dataset with customer transactions, I would approach the task systematically, using a combination of data pre-processing, exploratory analysis, and machine learning techniques. Below are the steps and tools I would use:

**Step 1: Understanding the Business Context**

* **Objective Clarification**: Understand the business objectives—why do we need to segment customers? Is it for targeted marketing, improving customer retention, or identifying high-value customers? Who are the customers?

**Step 2: Data Exploration and Pre-processing**

* **Data Collection**: Ensure that all relevant data is available, including transaction data (e.g., purchase amount, date, product categories), customer demographic data (e.g., age, location), and any other behavioural data.
* **Data Cleaning**: Handle missing values, correct inconsistencies, and remove duplicates. Standardize data where necessary.

**Step 3: Exploratory Data Analysis (EDA)**

* **Descriptive Statistics**: Calculate summary statistics (mean, median, etc.) for key features to understand the distribution of the data.
* **Data Visualization**: Use tools like Matplotlib, Seaborn, or PowerBI to visualize data distributions, correlations, and trends. For example:
  + **Histograms** for frequency and monetary value distributions.
  + **Box plots** to identify outliers.
  + **Heatmaps** to visualize correlations between features.
* **Customer Segmentation Hypotheses**: Based on the EDA, form initial hypotheses about possible customer segments. For example, frequent high-spenders might represent a distinct segment.

**Step 4: Customer Segmentation**

1. **Segmentation Techniques**:
   1. **RFM Analysis**: Use Recent, Frequency, and Monetary value to create customer segments based on purchasing behaviour. This method is simple yet effective for identifying high-value customers.
   2. **Clustering Algorithms**:
      1. **K-Means Clustering**: A commonly used unsupervised learning technique for customer segmentation. Choose the number of clusters (K) based on the elbow method or silhouette score.
      2. **Hierarchical Clustering**: Useful if the number of segments is unknown or if you want to understand the relationships between segments.
   3. **Demographic Segmentation**: Group customers based on demographic data such as age, income, or location. This can be combined with behavioural data for more detailed segments.
2. **Dimensionality Reduction**: If the dataset has a large number of features, use PCA (Principal Component Analysis)

**Step 5: Visualization**

* **Visualization**: Use 2D or 3D plots to visualize the clusters. Tools like Tableau, Power BI, or Python libraries (e.g., Seaborn, Plotly) can help in presenting these segments effectively.

**Step 6: Actionable Insights and Implementation**

* **Marketing Strategies**: Develop targeted marketing strategies for each segment, such as personalized offers for high-value customers or re-engagement campaigns for dormant customers.
* **Resource Allocation**: Allocate resources based on the potential value of each segment.
* **Monitoring and Iteration**: Implement a system to monitor segment behaviour over time. Adjust strategies based on changes in customer behaviour or business objectives.

**Tools Used:**

* **Python (Pandas, Scikit-Learn, Matplotlib, Seaborn)**: For data pre-processing, analysis, and clustering.
* **SQL**: For data extraction and manipulation from databases.
* **Tableau or Power BI**: For interactive data visualization and dashboard creation.
* **Jupyter Notebook**: For documenting the analysis process and presenting results.

**Conclusion:**

The process involves understanding the business needs, preparing the data, applying suitable segmentation techniques, validating the results, and deriving actionable insights. This systematic approach ensures that the segmentation is both meaningful and useful for driving business decisions.

**Question 10:**

Question: A company wants to launch a new product and has collected survey data on

customer preferences. How would you use this data to help the company make an informed

decision? Outline your approach.

**Approach:**

To help a company make an informed decision about launching a new product using survey data on customer preferences, I would follow a structured approach that involves data analysis, visualization, and deriving actionable insights. Here's how I would approach this task:

**Step 1: Understand the Objectives**

* **Objective Clarification**: Clearly define the company's goals for the new product launch. Is the goal to identify potential market segments, assess overall demand, or determine product features that appeal most to customers?
* **Key Questions**: Identify the key questions that the analysis should answer, such as:
  1. What are the most desired features?
  2. Which customer segments show the highest interest in the product?
  3. What price points are customers willing to pay?

**Step 2: Data Exploration and Pre-processing**

* **Data Collection**: Ensure that all relevant survey data is collected and available, including demographic information, customer preferences, willingness to pay, and other relevant factors.
* **Data Cleaning**: Handle missing data, remove outliers, and correct any inconsistencies in the survey responses. This ensures that the analysis is based on accurate and reliable data.
* **Data Transformation**: Convert categorical variables (e.g., customer preferences) into numerical form if needed, and create derived features such as average ratings for different product features.

**Step 3: Exploratory Data Analysis (EDA)**

* **Descriptive Statistics**: Calculate summary statistics to understand the overall distribution of responses. For example, calculate the mean, median, and mode for customer ratings of different features.
* **Segmentation**: Analyse customer demographics (age, income, etc.) to identify potential segments that show different preferences.
* **Feature Analysis**: Evaluate the importance of different features using frequency counts, averages, or scores. Identify which features are most and least liked by customers.

**Step 4: Data Visualization**

* **Preference Visualization**:
  + **Bar Charts**: Use bar charts to visualize the popularity of different product features or preferences among customers.
  + **Heatmaps**: Visualize the correlation between different features or preferences using heatmaps to identify patterns.
* **Demographic Analysis**:
  + **Pie Charts/Histograms**: Show the distribution of customer demographics such as age groups, income levels, or geographic regions.
  + **Box Plots**: Compare customer preferences across different demographic segments.
* **Willingness to Pay**:
  + **Scatter Plots**: Visualize the relationship between customer willingness to pay and their preference for certain features.
  + **Distribution Plots**: Show the distribution of price points that customers are willing to pay for the product.

**Step 5: Advanced Analysis**

* **Conjoint Analysis**: If the survey includes data on customer trade-offs between different product features, use conjoint analysis to determine the most valued features and optimal price points.
* **Sentiment Analysis**: If the survey contains open-ended responses, perform sentiment analysis to gauge overall customer sentiment towards the product and its features.
* **Cluster Analysis**: Perform clustering to identify distinct customer segments with similar preferences. This can help in tailoring the product to different market segments.

**Step 6: Deriving Insights and Recommendations**

* **Feature Prioritization**: Based on the analysis, prioritize the features that are most important to customers. For example, if the majority of customers highly value a particular feature, ensure it is included in the product.
* **Market Segmentation**: Identify key customer segments that show strong interest in the product. Tailor marketing strategies and product features to these segments.
* **Pricing Strategy**: Based on the willingness-to-pay analysis, recommend a pricing strategy that aligns with customer expectations while maximizing revenue.
* **Go/No-Go Decision**: Provide a data-driven recommendation on whether to proceed with the product launch. If the data suggests strong customer interest and alignment with the company’s goals, recommend moving forward. If not, suggest areas for product improvement or potential market repositioning.

**Step 7: Communicating Results**

* **Dashboard/Reports**: Create an interactive dashboard or report summarizing the key findings, using tools like Tableau, Power BI, or even Excel for straightforward visualizations.
* **Presentation**: Present the insights to stakeholders in a clear and concise manner, focusing on how the data supports the recommendations. Use visual aids like charts, graphs, and infographics to make the data easily understandable.

**Tools Used:**

* **Python (Pandas, Matplotlib, Seaborn)**: For data analysis and visualization.
* **Tableau/Power BI**: For creating interactive dashboards and visualizations.
* **Excel**: For quick calculations and simple visualizations.
* **Text Analysis Tools (e.g., NLTK,)**: For sentiment analysis on open-ended survey responses.

**Conclusion:**

This structured approach allows the company to make an informed decision based on customer preferences, ensuring that the product launch is aligned with market demand and customer expectations. By analysing the survey data thoroughly, the company can optimize the product features, pricing, and marketing strategies to maximize the chances of a successful launch.