1) Dimensionality Reduction Using Correlation and Heatmaps on Housing Dataset

Dataset Description:
The housing dataset contains various attributes related to residential properties, including features such as square footage, number of bedrooms and bathrooms, location, and sale prices.

Tasks:

Data Loading and Preprocessing:

- Load the housing dataset into your preferred data analysis environment (e.g., Python with Pandas).
- Perform necessary preprocessing steps such as handling missing values, encoding categorical variables, and scaling numerical features if needed.

Correlation Analysis:

- Compute the correlation matrix for the features in the dataset.
- Visualize the correlation matrix using a heatmap to identify highly correlated features.
- Interpret the heatmap to understand the strength and direction of correlations between different pairs of features.

Dimensionality Reduction Based on Correlation:

- Identify pairs of features with high correlation coefficients (e.g., correlation coefficient > 0.7 or < -0.7).
- Select one feature from each highly correlated pair for removal to reduce dimensionality.
- Discuss the rationale behind selecting specific features for removal and potential implications for model performance and interpretability.

Comparative Analysis of Dimensionality Reduction Techniques on Housing Datasets: Original vs. Kaggle Acquisition:

You have successfully applied dimensionality reduction techniques based on correlation analysis and heatmaps to a housing dataset in your previous lab exercise. Now, you have acquired a larger housing dataset from Kaggle (https://www.kaggle.com/c/santander-customer-satisfaction/data?select=train.csv), containing additional features and a larger sample size.

Compare the results obtained from the dimensionality reduction techniques applied to the original housing dataset and the new, larger dataset obtained from Kaggle. Discuss any differences or similarities observed in terms of:

- The nature and strength of correlations between features.

- The effectiveness of dimensionality reduction in improving model performance and interpretability.
- Any challenges or limitations encountered when applying dimensionality reduction techniques to the larger dataset.

Provide insights into how the characteristics of the new dataset may influence the selection and application of dimensionality reduction methods and discuss potential strategies for addressing these challenges.

2) Tableau:

Data: Student Performance in Exams
https://www.kaggle.com/spscientist/students-performance-in-exams

- Study the data and come up with 3 different ways of visualizing the data using Tableau.
- Create a Tableau dashboard.
- Summarize what you learnt from your analysis.
- Study the Titanic data set and come up with a dashboard that shows the number of survivals among males and females based on the age (age should be converted to bins of size 10).

3) Excel:

- Study the SalesManagers data set.
- Using Excel pivot tables, find the best sales manager per region.
- Using Excel pivot tables, find out who the best sales manager is among the whole list.
- Which item bring the highest revenues (sales) for the company?

4) Visualization on "**churn.csv**" dataset:

In the following questions, you need to use Python libraries to get as close as possible to the info shown in the figure. You don't need to get it exactly the same as these charts are based on SPSS, which was used in "Discovering Knowledge in Data" book.

a) Investigate using bar chart the churn variable and print the percentages (Fig. 1). Comment on that (any insights?).

Value	Proportion	%	Count
False		85.51	2850
True		14.49	483

Fig. 1

b) Investigate the "International Plan" variable using stacked bar chart (Fig. 2) and normalized stacked bar charts (Fig. 3). What are your conclusions?

Fig. 2

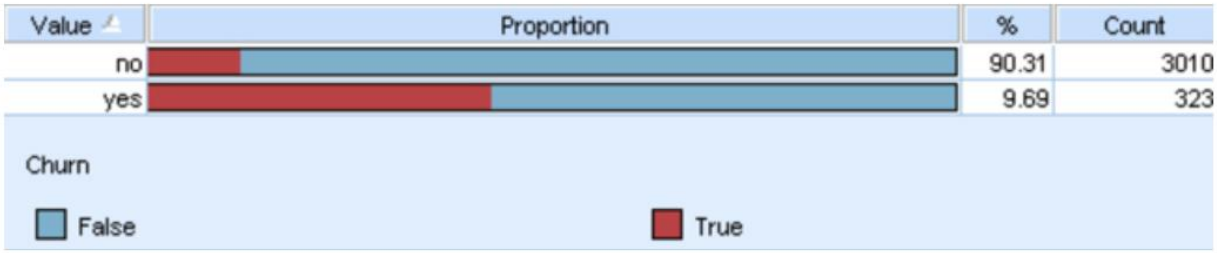

Fig. 3

c) Generate the contingency table (Fig. 4):

		International Plan		
		No	Yes	Total
Churn	False	Count 2664 Col % 88.5%	Count 186 Col % 57.6%	Count 2850 Col % 85.5%
	True	Count 346 Col % 11.5%	Count 137 Col % 42.4%	Count 483 Col % 14.5%
	Total	3010	323	3333

Fig. 4

d) Investigate the “International Plan” variable as shown in Fig. 5 and Fig. 6.

Fig. 5

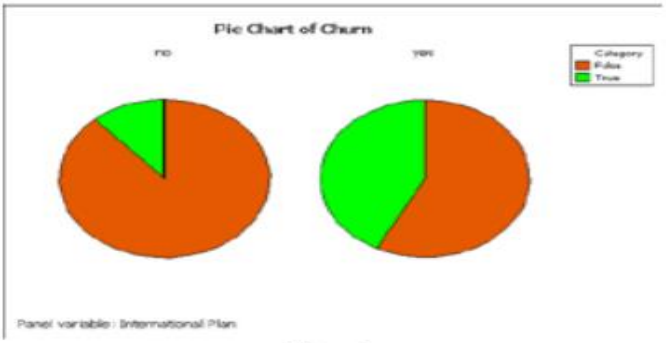

Fig. 6

e) Explore 5 of the numeric variables and get some descriptive statistics about these variables as shown in the Fig. 7. Any insights/ observations you can get?

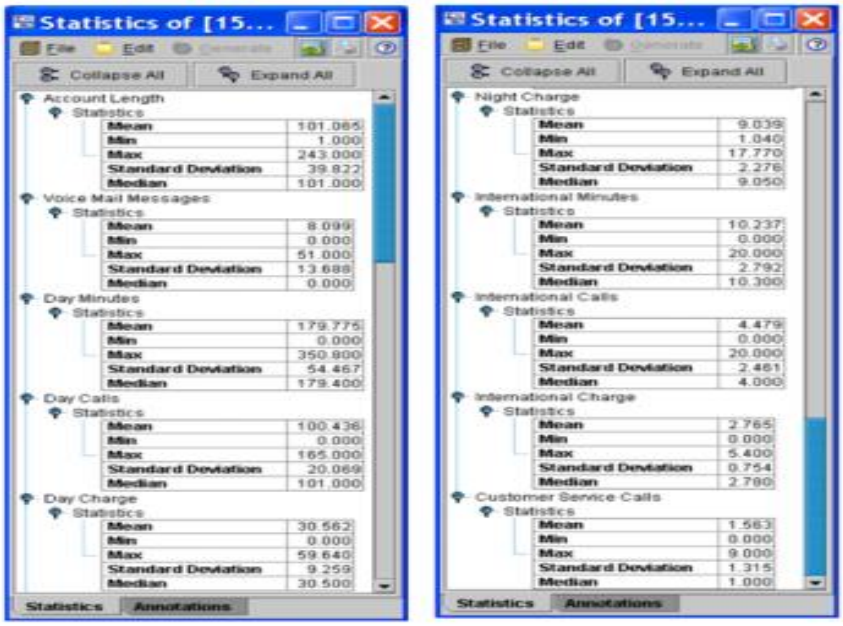

Fig. 7

f) Get a histogram with churn and without churn (Fig. 8) for the customers service call variable, what do you conclude?

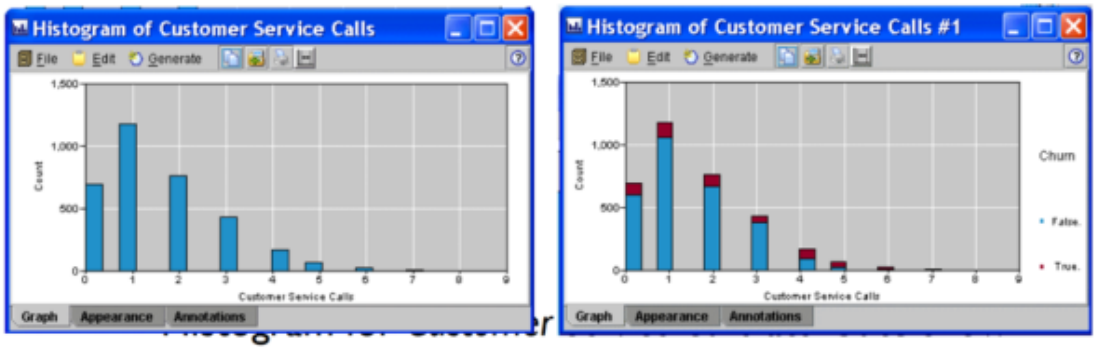

Fig. 8

g) Analyze the multivariate relationships between Day Minutes and Evening Minutes:

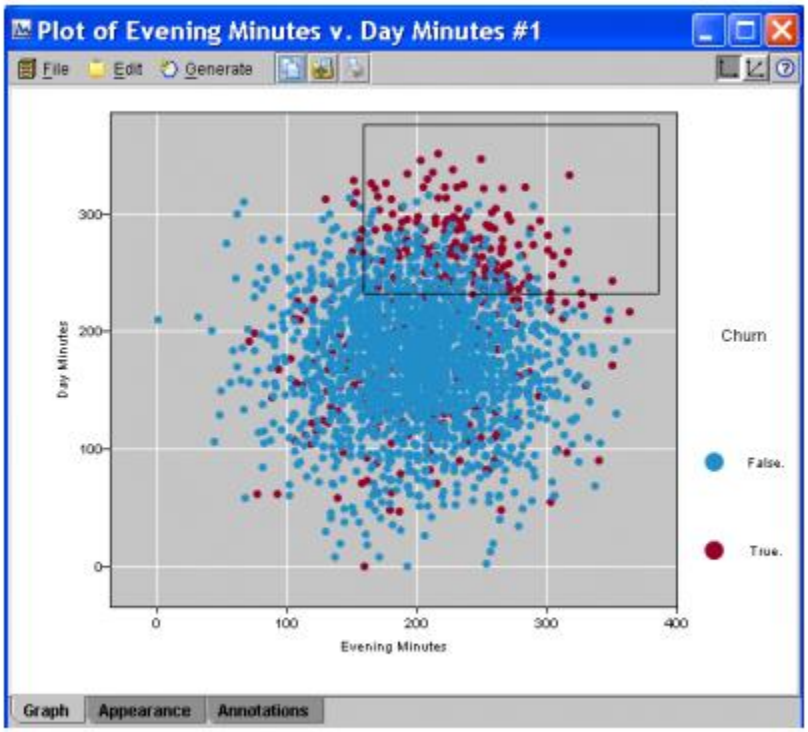

Fig. 9

h) Bin the Customer Service calls variable into Low (< 4 calls) and high (>= 4) and get the churn rate for Low and High customers service calls.

i) Get the correlation plot for Day Minutes, Day Charge, Night Minutes, Night Charge variables (Fig. 10).

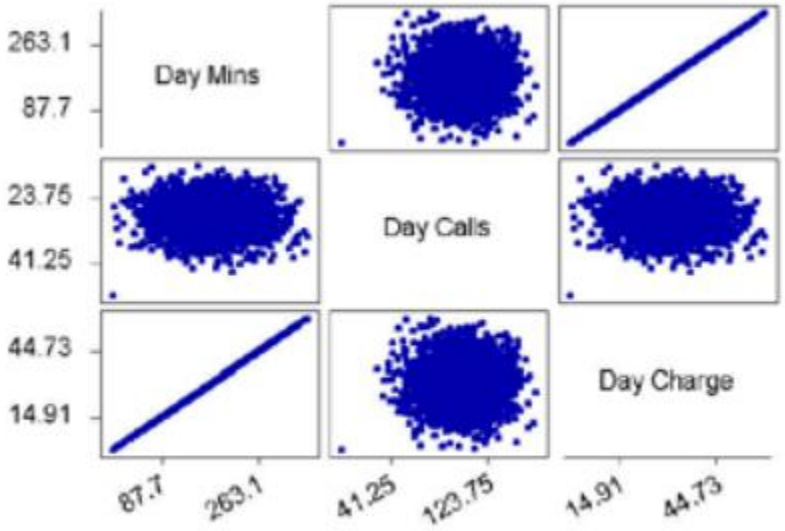

Fig. 10