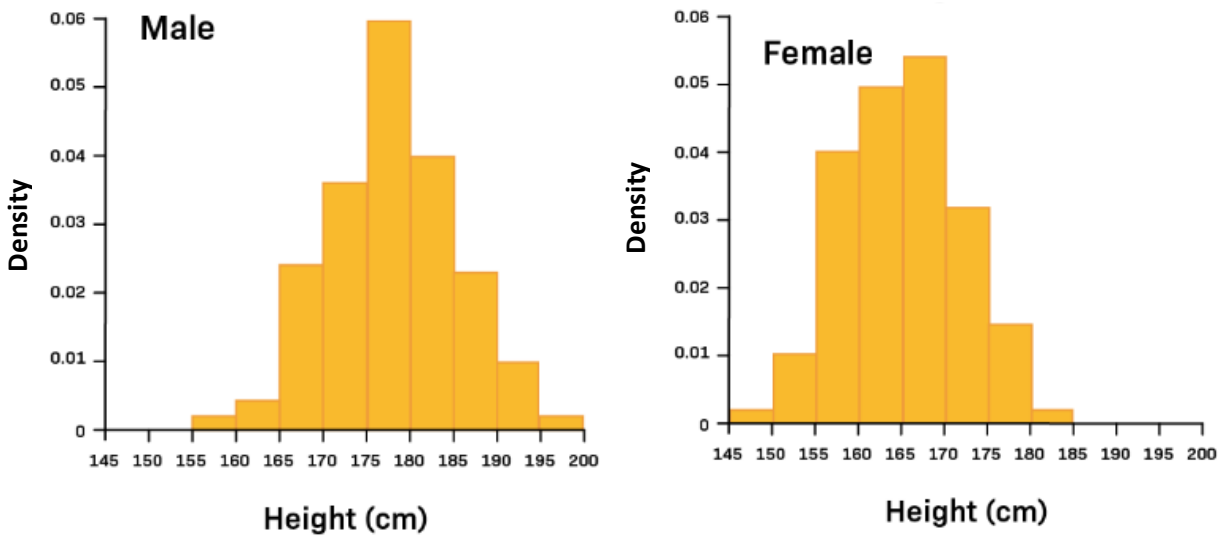
MSD522: Data Mining for Business
Mid-Semester Examination
Monsoon Session – 2022-23

Total Marks: 28 **Time: 2 Hours**

1. Which type of data visualization chart would you use to understand the relationship between two continuous variables? Explain with the help of an example. **[2 marks]**

2. What do you infer by comparing the two graphs below: **[2 marks]**



3. A company is planning to land the operational system data in the data warehouse for analytical purposes. As a data operations consultant, which database schema would you recommend to the company considering the sample dataset given below, and why? Also, develop the recommended schema using the below sample dataset. **[8 marks]**

Order ID	Order Profit	Order Quantity	Item Name	Item Color	Warehouse State	Warehouse City	Manager Name	Employee Name
101	Rs. 100000	1	Sedan	Blue	Texas	Houston	John	Jane
102	Rs. 200000	2	Sedan	Blue	Florida	Orlando	Phil	Joe
103	Rs. 200000	2	Sedan	Blue	Texas	Houston	John	Jill
104	Rs. 400000	2	SUV	Brown	Texas	Houston	John	Jane
105	Rs. 800000	4	SUV	Brown	Florida	Orlando	Phil	Jill

Employee Gender	Employee Office	Employee phone	Month	Year	Quarter	Customer Name	Customer Address	Customer Phone
F	Utah	xxx	May	2022	Q1	Bill	123 quarter	xxx
M	Texas	xxx	May	2022	Q1	Ben	456 quarter	xxx
F	Texas	xxx	May	2022	Q1	Ben	789 quarter	xxx
F	Utah	xxx	May	2022	Q1	Bill	xyz quarter	xxx
F	Texas	xxx	May	2022	Q1	Ben	abc quarter	xxx

4. A bank, which specializes in giving housing loan, wants to automate the loan eligibility process based on the details supplied by the customers at the time of filling out loan

application. The bank manager has reached out to your team to help them identify the segments of customers who are eligible for loan, so that the bank personnel can specifically target those customers. Using the below details supplied by the customers, answer following questions: **[10 marks]**

a. Is it a supervised learning problem or unsupervised learning problem? Why?
b. For the given data below, what data pre-processing steps would you be following? Also, show the implementation of those steps on the given data.
c. List down and implement the feature engineering processes wherever required.

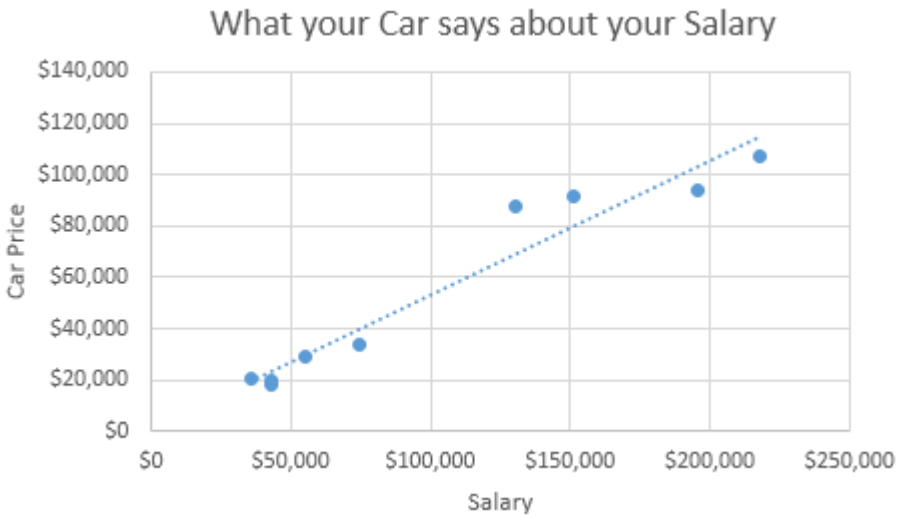
Loan_ID	Gender	Married	No. of Dependents	Education	Self_Employed	Applicant Income	Co-applicant Income	Loan Amount	Loan_Amount_Term (Days)	Credit_History	Loan_Status	Property_Area	Total Income
LP001002	Male	No	0	Graduate	No	5849	0		360	Good	Y	Urban	$5849.0
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	Good	N	Rural	$6091.0
LP001005	Male	Yes	0	Post Graduate	Yes	3000	0	66	360	Good	Y	Urban	$3000.0
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	Good	Y	Urban	$4941.0
LP001008	Male	No	0	Graduate	No	6000	0	141	360	Good	Y	Urban	$6000.0
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	Good	Y	Urban	$9613.0
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	Good	Y	Urban	$3849.0
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	Bad	N	Semiurba	$5540.0
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	Good	Y	Urban	$5532.0
LP001020	Male	Yes	1	PhD	No	12841	10968	349	360	Good	N	Semiurba	$23809.0
LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	Good	Y	Urban	$3900.0
LP001027	Male	Yes	2	Graduate		2500	1840	109	360	Good	Y	Urban	$4340.0
LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	Good	Y	Urban	$11179.0
LP001029	Male	No	0	Graduate	No	1853	2840	114	360	Good	N	Rural	$4693.0
LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	Good	Y	Urban	$2385.0
LP001032	Male	No	0	Graduate	No	4950	0	125	360	Good	Y	Urban	$4950.0
LP001034	Male	No	1	Not Graduate	No	3596	0	100	240		Y	Urban	$3596.0
LP001036	Female	No	0	Graduate	No	3510	0	76	360	Bad	N	Urban	$3510.0
LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	Good	N	Rural	$4887.0
LP001041	Male	Yes	0	Graduate		2600	3500	115		Good	Y	Urban	$6100.0
LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	Bad	N	Urban	$7660.0

5. Should we drop highly correlated variables before performing Principal Component Analysis (PCA)? **[3 marks]**

6. What will happen if eigenvalues are roughly equal across principal components in PCA? **[3 marks]**

Model Answer

1. To understand the relationship between two continuous variables, we can use Scatter plot. Scatter plots are extremely useful in identifying the patterns in data, association between two variables. For example, if we want to know the patterns between salary of the individual and the price of the car bought by the individual, we can plot this using scatter plot like below.



As indicated in the above chart, individuals with higher salary prefer costly cars may be with enhanced features. One thing should be kept in mind that scatter plot does not indicate causality but association between two variables.

2. The two graphs are the histograms for the heights of male (aka Fig 1) and female (aka Fig 2) respectively. Histogram is a graph that helps us to understand the shape of the data distribution of a numeric variable. It is helpful to identify the most common values and least common values in a dataset. Each bar represents a bin. Y-axis is the proportion of occurrences in each bin. Male height distribution is almost similar to normal distribution where the mean male height lies between 175-180 cm. Moreover, mean female height is ranging between 165-170 cm and the data distribution is not symmetrical.

3. Given the data, I would suggest Star schema. Star schema are generally denormalized and they support OLAP cubes which makes processing of the query faster and easier. Data aggregation operations can easily be applied on the data as data is stored in OLAP cubes. The data model as per star schema is given below:

4. a) It is a supervised learning problem as the bank wants to develop a model which identifies the customers who are eligible for loan. Hence, the target variable is loan status.

   b) Treat missing values in self-employed and credit history column. Create a new category “unknown” for missing value. Impute the most frequent value in loan amount term (days) feature missing cells. There is an outlier in Applicant income column and Total income column. I would keep this outlier as it is a natural outlier. It is a natural outlier because the education of the applicant is PhD and being one of highest qualifications, the applicant is getting paid the highest amount. Remove $ sign from total income feature column. Remove + sign from 3+ in No. of dependents.

   c) Drop loan ID feature from the dataset. Generate one hot encoded features for Self-employed, credit history and property area. Label encode gender, married, education and loan status. Standardize the Applicant income, Co-applicant income, loan amount and total income features.

5. No, that is the ultimate objective of PCA.PCA loads all highly correlated variables on one principal component.

6. In such case when all eigenvalues are roughly equal, PCA wouldn't be able to explicitly select the principal components. The other approach of selecting PCs could be checking for the cumulative variance that is explained by obtained PCs. Cumulative variance can be obtained by arranging the eigen values in decreasing order and summing the explained variance. Model developer may set a threshold of explained variance and then select the number of Principal Components that generate a cumulative sum of explained variance higher than the threshold variance. But, this approach has flaws as the threshold is subjective.