Assignment Based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A1. The following are the inferences drawn from the analysis on categorical variables-

- 32% of the bike booking were happening in fall with a median of over 5000 booking. This was followed by summer & winter with 27% & 25% of total booking. Season can be a good predictor for our model.
- Almost 10% of the bike booking were happening in the months may, jun, jul, aug & sept with a median of over 4000 booking per month. So mnth is also a good predictor for our model.
- Approximately 68.4 percent of bookings are done on a weekday and thus can be a good predictor of our model.
- As we can see 97.6 percent sales happening on non holidays, which means we cannot rely on this data as it is clearly biased. So we will skip holiday from our model.
- There is some trend here. We can see 68.6 percent of sale happens on Clear-PartlyCloudly situation. While 30.2 percent comes from Mist-Cloudy days. So there is a trend. Also the median lies around 5000 for Clear-PartlyCloudly situation. Thus we will consider weathersit is our model

Q2. Why is it important to use drop_first=True during dummy variable creation?

A2. It is important to use drop_first=True during dummy variable creation since it removes the addition of redundant feature to the model. Let's take an example-

For example there is a column that contains gender details as Male or Female, now when we create dummy variable out of it, if we don't do drop_first=True, it will create 2 columns Male and set 1 for users who are male and 0 otherwise. Similarly, the other column will have same case for females as well but if we think, one column is redundant to identify the other column. That is we can identify male or female from 0 or 1(in only one column). Similarly for other variables where the count of categories > 2, we can always consider dropping a single column, since it can be identified on the basis of other categories binary values.

Q3 **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

A3. As the target variable is cnt, the one which has the highest correlation with the target variable is atemp.

Q4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

A4. The assumptions were validated using the following steps -

a. Validated if the error terms follow the normal distribution.
b. Ensured that there is no multicollinearity among the predictor variables by seeing the VIF values.
c. Checking homoscedasticity of error terms .

Q5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

A5. According to final model, the top 3 features contributing significantly towards explaining the demand of shared bikes are-

a. Temperature
b. Year
c. Weathersit_Light Snow (Negative)

We can infer that users prefer to drive the bike in a moderate temperature and thus the sales get impacted in at those time.
Also with year we can say that once the situation is back to normal after COVID, the company will stop seeing the declines and will be a normal function.
Also users have a negative impact due to LightSnow-lightRain-Thunderstorm, that is in this season people don't prefer to have a bike.

General Subjective Questions

Q1. **Explain the linear regression algorithm in detail.**

A1. Linear regression is a Machine Learning model that kind of measures the relationship between one or more predictor and one outcome variables. It is used for predictive analysis and modelling.
There are 2 ways we classify linear regression algorithm into-

a. Simple linear Regression
b. Multiple Linear regression

Simple Linear Regression

Simple linear regression is a regression model where we have just one predictor or independent variable. The equation to determine the model represents something like: $Y = mx + c$, which is basically a straight line.

Multiple linear regression

MLR attempts to model the relationship between 2 or more predictor variables and a response variable by fitting a linear equation to observe data. Every value of predictor variable x is associated with a value of dependent variable y.

In both the models, we calculate the coefficient for each of the independent variable x, which resembles the impact of that particular variable in defining the whole model.

We split the data we have into 2 parts-

a. Train set
b. Test set

Both of these come from the data we have. This is done to know how reliable our model is.

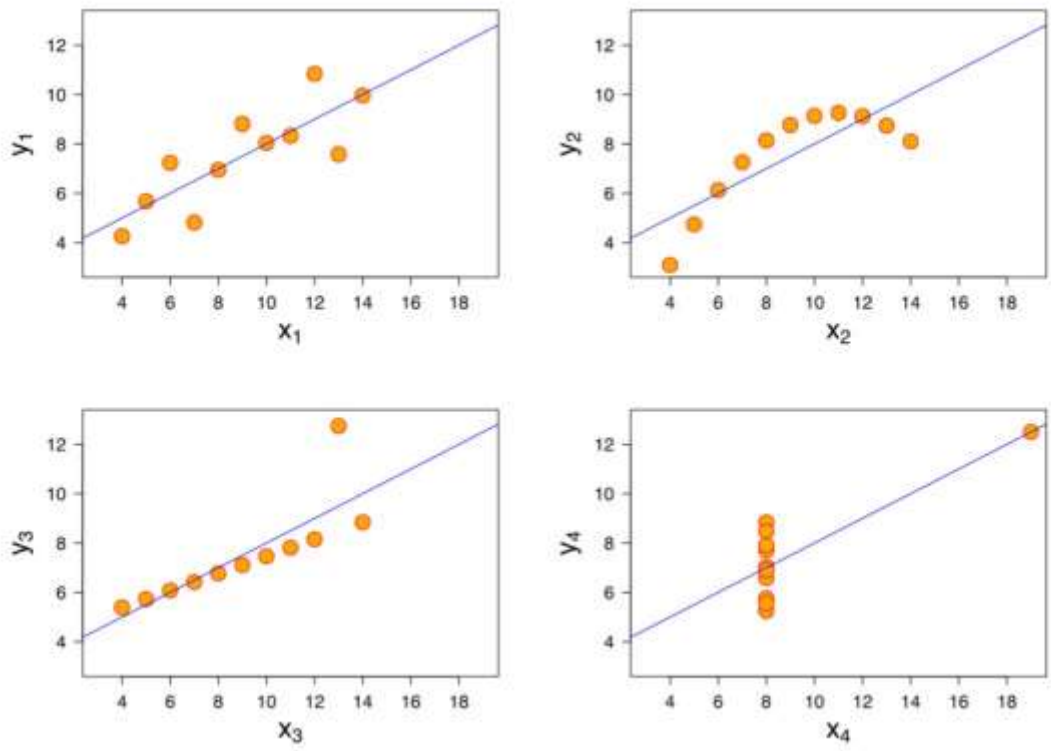
Using the coefficients we found using the training set, we predict the values of the test set.

Now to evaluate the performance of the model, we see the r-square and adjusted r-square values. Higher the r-squared better the model. RMSE is the standard deviation of the residuals. Now residuals are a measure of how far form the line are the real data points.

Q2. **Explain the Anscombe's quartet in detail.**

A2. Anscombe's quartet comprises four datasets, each containing eleven (x,y) pairs. The essential thing to know about these datasets is that they have the same centric values, like mean median.. But things change completely when they are visualized. Each graph tells a different story irrespective of their similar summary statistics.
The summary statistics show that the means and the variances were identical for x and y across the groups :
Mean of x is 9 and mean of y is 7.50 for each dataset.
Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

Now once we plot them, we see a different story altogether.



As we see in the diagram-

- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This means the stats calculation are not enough to determine the data, we need to visualize it as well.

Q3. **What is Pearson's R?**

Correlation between data is a measure of how well they are related to each other. The most commonly used measure of correlation in stats is the Pearson Correlation.

The Pearson correlation coefficient, r, can take on values between -1 and 1. Higher the absolute value of the coefficient stronger the linear relationship between the two variables. The sign of r corresponds to the way the variable effects the other. If r is positive, then as one variable increases, the other tends to increase. If r is negative, then as one variable increases, the other tends to decrease. A perfect linear relationship (r=-1 or r=1) means that one of the variables can be perfectly explained by a linear function of the other

Q4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

A4. Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying values. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. Scaling is performed to bring the values of numerical variables in a certain scale so that the model is able to perform better and not go on higher values. This kind of normalizes the data to a scale.
The difference between Normalized scaling and Standardised scaling

a. Normalization(Min-Max)
   This rescales the variable values with distribution between 0 and 1.
b. Standardization
   This rescales a feature value so that it has distribution with 0 mean value and variance equal to 1.

Q5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? Why is it ?**

A5. VIF stands for Variance inflation factor.
In LR, collinearity can make the coefficient unstable which will cause coefficients to be less reliable and p-value will be more.
VIF is dependent on the r2 which is the correlation among 2 variables.
If the R^2 is high which means there is a high correlation among variables the VIF will be higher. The formula is VIF = 1/(1-R2). So which means if we have perfect correlation among variables the VIF value will reach infinity. The principle to follow is to remove the features having VIF > 5.

Q6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression**.

A6. Q-Q plot also known as, Quantile-Quantile plot is a graphical tool to see if a set of data came from some distribution such as Normal, Exponential or uniform. Also it helps to determine if 2 data sets came from populations with common distribution. Now this is helpful in Linear regression. It helps when we have training and test data set received separately and using the Q-Q plot we can confirm if the data sets are from population with same distribution.

Advantages-

a. It is useful with sample sizes.
b. Even the scale shift or presence of outliers can be detected from this plot.

It is used to check 2 data sets came from population with common distribution, and have a common scale.

A Q-Q plot is basically a plot of the first data set against the quantiles of the second data set.