Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans. We have done the analysis of categorical column using box plot and below are the inferences.

- Spring season has the least booking and fall season seemed to be highest booking. However, mean of fall and summer and winter is very close compared with fall's rental
- In terms of year, business looks to be growing as we can observe that 2019 has very high rentals than 2018.
- When plotted for months in a year it was seen that Sep had highest booking and Jan being the lowest
- Weekday or not didn't have much impact on the bike rentals, same goes with working day or not.
- Weather had a clear impact on the rentals. The highest number of bookings were on the clear/ less cloudy days. The booking seems to be dropping as weather changes from clear to rainy to heavy rainy. There are no bookings on high rainy days as expected.

2. Why is it important to use drop_first=True during dummy variable creation?
Ans. Drop_first=True is important to use since it helps in reducing 1 variable from data while creating dummy variables. If we have 3 distinct values of a categorical variable we need only 2 dummy variables. Hence, to generalize, categorical variable with ***n*** unique values needs only ***n-1*** variables .

We used these feature in the assignment as below syntax:
season_df=pd.get_dummies(bike["season"],drop_first=True)
Default value of drop_first is False

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans. 'temp' and 'atemp' variables have the highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans. We have validated the assumptions of linear regression using below tests.

- **Homoscedasticity** : We observed that there are no visible pattern in errors
- **Normality in error terms**: We plotted the errors distribution plot and found that they are normally distributed
- **Linearity**: Independent variables are linearly related with target variable
- **Multicollinearity**: We checked for multicollinearity using heat-map and observed very low correlation
- **Auto-correlation**: Using Durbin watson test we observed that there is no correlation between residuals

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans. Based on the coefficients, the top 3 features explaing the demands are:
a. Temp (positive relation)
b. Year (positive relation)

c. Light_Snow_Rain weather (negative relation)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear Regression is the statistical methodology used for predictive analysis. It works on the idea that several independent predictor variables which are linearly realated to the target variable can predict the target variable (dependant variable). In this, we examine which predictors are highly impacting the prediction of the target variable indicated by the magnitude and sign of the relation.

The simplest form of linear regression equation is single variate , denoted by

y=mX + C

Where y is the target variable ,
m is the slope or coefficient of X, it shows the impact of X on y by its magnitude and sign,
X is the independent variable,
C is the constant or the intercept in the line. If x=0 , then y=c

Based on the sign of coefficient the linear equation can be positive or negative.
Positive linear relation means with increase in X , y will increase, and negative means with increase in X, y decreases

Linear regression can be simple linear regression or multiple linear regression.
Simple regression has 1 independant variable and its equation is given as
Y=mX+c

And multiple regression has more than 1 independent variable and its equation is given as:
y=X0 + a*X1 + b*X2 + c*X3 ……

Where X0 is the intercept and X1 X2 are the predictors
and a,b,c are the coefficients of the predictors.

There are several assumptions for linear regression:

a. Multi collinearity:

The linear regression model assumes that there is very little or no multicollinearity between the predictor variables. It means the predictor variables should be independent and they should be defined using other predictors. It can be checked using VIF and to treat it we can remove the variables with high VIF (>10)

b. Auto-correlation:

Model assumes that there should not be auto correlation between residuals. It can be checked using Durbin watson method and its value ~2 is desirable for non-autocorrelation

c. Linear relation

There should be linear relationship between each predictor and target variable

d. Normality of error terms:

The error terms should be normally distributed

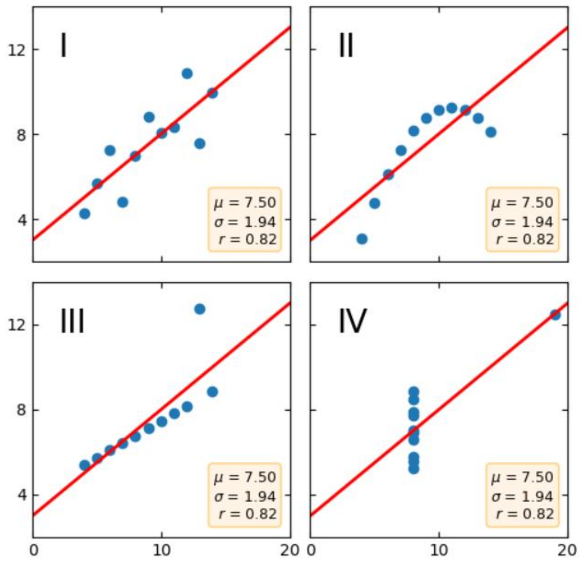
e. Homoscedasticity:

There should not be any pattern in the residuals

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet comprises of 4 datasets with nearly same statistical description like count, mean, variance etc. However, when visualized it is observed that all 4 data is totally different from each other. I also shows the effect of outliers on the numerical calculations.

These datasets and its results were constructed by the statistician Francis Anscombe to show the importance of visualization over the numerical calculations. Below is the graphs of the 4 datsets.



(Image source: https://matplotlib.org/stable/gallery/specialty_plots/anscombe.html)

Each dataset has 11 data points (x,y)
And their mean, variance , linear regression line , R2 all are equal or very close to each other. However, when plotted, shows the totally different result.

3. What is Pearson's R?

Ans. Pearson correlation coefficient, also known as Pearson's R is the measure of correlation between variables. If one variable increases with another variable the coefficient will be positive , however if 1 variable go up and another goes down the coefficient will be negative.

Its value lies between -1 and 1. Value 0 denotes that there is no correlation between the variables and there is no relationship between them. 1 denotes there is perfect relation between them

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

Ans. Scaling is done to standardize the values of continuous variables in the fixed range. This is a part data pre processing. The variables with very high values are brought in the range with other variables so that when coefficients of the model is derived they are comparable and easy to analyse. Also, if the variable values are in the same range their coefficients can be observed to know which predictor has high impact on the model equation.

Normalized scaling rescales the values to be in the range [0,1] or [-1,1]however Standardized scaling doesn't have any range but it will have the data to have the standard deviation of 1 and mean 0.

Normalized scaling or min max scaling is calculates as :

X_new = (X - X_min)/(X_max - X_min)

Standardized Scaling is calculated as :

X_new = (X- mean) /Std

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

Ans. If there is perfect correlation between 2 or more variables then VIF will be infinity. Since VIF is given as 1/(1-R2), hence if R2 becomes 1 then VIF becomes infinite.
Which means that if 1 variable can be perfectly explained by any 1 or more variable then R2 will become 1.
In this case we need to drop the variable causing VIF infinity.
In the model we created, we had dropped the variable atemp because it was highly correlated with the variable temp, otherwise it would have shown us the VIF as infinity. Since it was perfectly explained and had linear relation with the variable *temp*.

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

Ans. Quantile- quantile plot is a graphical technique to plot a quantile of a datset against same quantile of another dataset. This is doone to ensure if the 2 datasets come from population with same distribution.

This is useful when we receive train and test data separately. So to ensure that they come from population with same distribution we plot Q-Q plot and if they are same then the plot will be a straight line y=x.
It is also useful to show if the residuals follow a normal distribution.