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? (3 marks)
   Answer: *In our final model (where we used RFE to identify relevant features), we have used below categorical variables*
   - *workingday: have a positive effect on the demand (~0.05). Rides are more on working days*
   - *season_spring: have a negative effect on the demand (~0.1)*
   - *season_winter: have a positive effect on the demand (~0.07)*
   - *mnth_dec: have negative effect on demand (~0.06)*
   - *mnth_jan: have negative effect on demand (~0.06)*
   - *mnth_nov: have negative effect on demand (~0.06)*
   - *mnth_sep: have positive effect on demand (~0.06)*
   - *weekday_saturday: have positive effect on demand (~0.06)*
   - *weathersit_mist: have negative effect on demand (~0.08)*
   - *weathersit_snow: have negative effect on demand (~0.3). It clearly has more impact than mist*
     Snow seems to have the highest impact on the demand in a negative way by a factor of 0.3.
     Working days see more demand than holidays
     Except September, other months show a decline in demand

     NOTE: the factors/coefficient assumes all other features have a constant impact

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   Answer: *If we do not use drop_first=True, then n dummy variables will be created for n values of a categorical variable. These n dummy variables are themselves correlated (multicollinearity). This would result in what is normally called as Dummy Variable Trap*
   For eg:- if we have a feature gender,
   which has male and female value. If we create 2 dummy variables then these 2 variables are correlated
   Multicolinearity would reduce model effectiveness. In Linear Regression one of the assumption is that there should no multi-colinearity

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Answer: *temp and atemp 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? (3 marks)
   Answer:
   Linearity – *Based on the pair plots we could see a linear relationship between the features and target variable*
   No multi-colinearity for features – *We have used VIF to check multi-colinearity. VIF<5 is considered a good threshold and all our features have VIF less than 5*
   No auto-correlation in residuals – *Durbin-Watson test can be used for this. Value close to 2 suggests that there is no auto correlation. We have got a value of 1.9*
   Error terms should be normally distributed – *We plotted (distplot) of the residuals for the train data and we could see normal distribution*
   Homoscedacity – *We have plotted a scatter plot of predicted train data vs residuals and we cannot find any obvious patterns which indicate homoscedacity*

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Answer:
   - weathersit_snow, year and temperature has highest impact on the demand
   - snow impacts demand negatively: People ride less when snowing
   - year impacts positively: Demand seems to be increasing with year (Time)
   - temperature impacts positively: Lesser temperature not favorable for demand

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   Answer: *Linear regression is a supervised machine learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.*
   Linear regression provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. They are of 2 types – Simple Linear Regression and Multiple Linear Regression

   <u>Simple Linear Regression</u>*: There is one dependent variable and 1 independent variable*
   *Formula: y(x) = p0 + p1 * x*
   where,

   - *y = output variable. Variable y represents the continuous value that the model tries to predict.*
   - *x = input variable. In machine learning, x is the feature, while it is termed the independent variable in statistics. Variable x represents the input information provided to the model at any given time.*
   - *p0 = y-axis intercept (or the bias term).*
   - *p1 = the regression coefficient or scale factor. In classical statistics, p1 is the equivalent of the slope of the best-fit straight line of the linear regression model.*

   <u>Multiple Linear Regression</u>*: There is one dependent variable and multiple independent variable*
   Formula: y(x) = p0 + p1x1 + p2x2 + … + p(n)x(n)

   For linear regression to hold for a model, the data should satisfy below assumptions

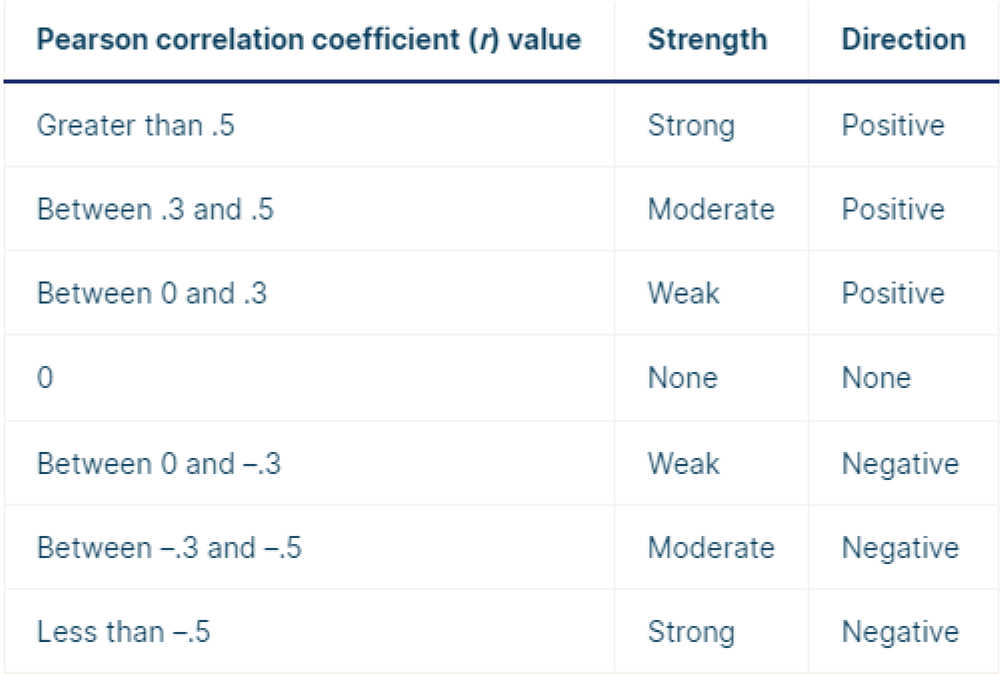
- ***Linear model**: the relationship between the independent and dependent variables should be linear.*
- ***No multi-colinearity in the data**: Independent variables (Features) should not be correlated*
- ***Homescedacity** of the residuals (Equal variance): Homoscedasticity means that they are roughly the same throughout, which means your residuals do not suddenly get larger or smaller*
- ***No autocorrelation in the residuals:** Residuals should not be autocorrelated*
- ***Errors terms/residuals are normally distributed:** Residuals should follow normal distribution.*

2. Explain the Anscombe's quartet in detail. (3 marks)
   Answer: *Anscombe's quartet comprises of four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.*
   This tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).
   For instance, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson's R? (3 marks)
   Answer: *The Pearson correlation coefficient (*r*) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.*
   This is also called as

- *Pearson's r*
- *Bivariate correlation*
- *Pearson product-moment correlation coefficient (PPMCC)*
- *The correlation coefficient*

This describes the strength and direction of the linear relationship between two quantitative variables.



Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

Formula:

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x}^2 - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y}^2 - (\sum \mathbf{y})^2]}}$$

Where,

- *r = Pearson Coefficient*
- *n= number of pairs of the stock*
- *∑xy = sum of products of the paired stocks*
- *∑x = sum of the x scores*
- *∑y= sum of the y scores*

- *∑x2 = sum of the squared x scores*
- *∑y2 = sum of the squared y scores*

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

Answer: *Scaling (Feature Scaling) is used 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 magnitudes or values or units. For eg:- we can have weight of people and height of people which are in different units.*

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. This is important for feature interpretation

Types of scaling –

***Normalized scaling**: Also called Min-Max scaling. This is used to transform features to be on a similar scale. The new point is calculated as:*

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

This scales the range to [0, 1] or sometimes [-1, 1]. Geometrically speaking, transformation squishes the n-dimensional data into an n-dimensional unit hypercube. Normalization is useful when there are no outliers as it cannot cope up with them

***Standardized scaling**: This is also called z-scale normalization. Here we subtract feature value from mean and divide by standard deviation. This is often called as Z-score.*

X_new = (X - mean)/Std

Geometrically speaking, it translates the data to the mean vector of original data to the origin and squishes or expands the points if std is 1 respectively. We can see that we are just changing mean and standard deviation to a *standard normal distribution which is still normal thus the shape of the distribution is not affected.*

Standardization does not get affected by outliers because there is no predefined range of transformed features.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Answer: *An infinite VIF would mean a perfect correlation between the variables. (Variables have perfect linear relationship with the other variables). This is a rare case*

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Answer: *Q-Q plot is short for quantile-quantile plot. This is a graphical technique to confirm if 2 data sets come from populations with a common distribution.*
   This is essentially a plot of the quantiles of the first dataset against the quantiles of the second dataset.
   A quantile means the percentage/fraction of points below the given value. Eg:- 0.4 or 40% would mean that 40% of the data would like below the value and the rest 60% will be above the value.
   In this method, we draw a 45-degree reference line. If 2 sets come from a population with same distribution, the points should fall approximately along this reference line. The greater the departure from the reference line, the greater the evidence that the 2 datasets belong 2 population with different distribution.
   In linear regression, we can use this technique to identify if the training and test data set are from population with same distributions. Also we can check if they have common scale, similar distribution shapes and similar tail behavior