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)

Ans: Categorical variables 'season', 'holiday', 'mnth', 'yr', 'weekday', 'workingday', 'weathersit' We can infer the following from these by analyzing boxplots of these variables vs target variable (cnt= count of total rental bikes):
season: fall and summer cover maximum variance followed by winter and spring i.e. more people are renting bikes in fall and summer.
mnth: The median starts increasing from January till July and then starts to decrease but the distribution across different months is almost same.
holiday: maximum variance is covered by non-holiday days i.e. more bikes are rented on non-holidays (working days).
yr(year): more bikes were rented in 2019 than 2018.
weekday: medians of weekdays lie at approximately equal levels.
workingday: more people rent bikes on working days.
weathersit: most bikes were booked on clear weather days followed by mist and then least in light/rain.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Ans: **drop_first=True** is **important to use**, as it helps in reducing the extra column created **during dummy variable creation**. Hence it reduces the correlations created among **dummy variables** and deletes the redundancy.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: **atemp and temp** has 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)

Ans: After creating the final model, we performed Residual Analysis to validate following assumptions:
1. Checked whether error terms are normally distributed with mean zero (not X, Y) by creating the histogram of error terms.
2. Checked whether error terms are independent of each other and error terms have constant variance (homoscedasticity) by creating plot between residuals and predicted values of y.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

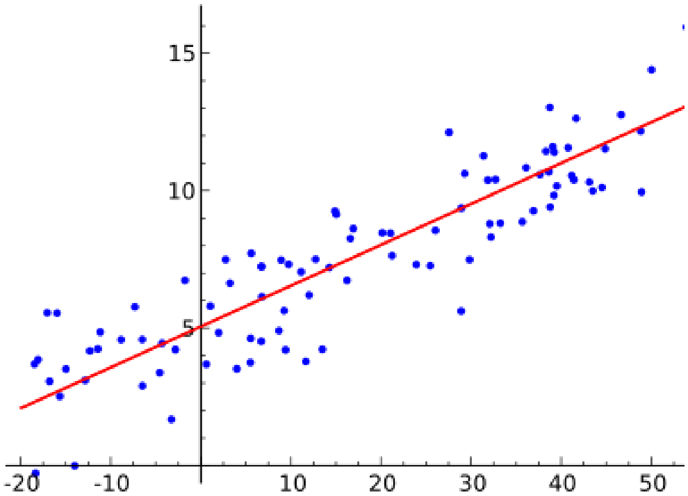
Ans: Based on the final model, top 3 features contributing significantly towards explaining the demand of the shared bikes are:
1) weathersit_light (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
2) year_2019(which is basically year)
3) season_spring

General Subjective Questions

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

- Ans: Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).

- It is a useful tool for predicting a quantitative response. Let's say we have a dataset which contains information about the relationship between 'number of hours studied' and 'marks obtained'. A no. of students have been observed and their hours of study along with their grades are recorded. This will be our training data. Our goal is to design a model that can predict the marks if the number of hours studied is provided. Using the training data, a regression line is obtained which will give minimum error. This linear equation is then used to apply for new data. That is, if we give the number of hours studied by a student as an input, our model should be able to predict their mark with minimum error.



- **Hypothesis function for Linear Regression:**

$$y = \theta_1 + \theta_2.x$$

- While training the model we are given:

  x: input training data (univariate – one input variable(parameter))
  y: labels to data (supervised learning)
- When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.

- **θ1:** intercept

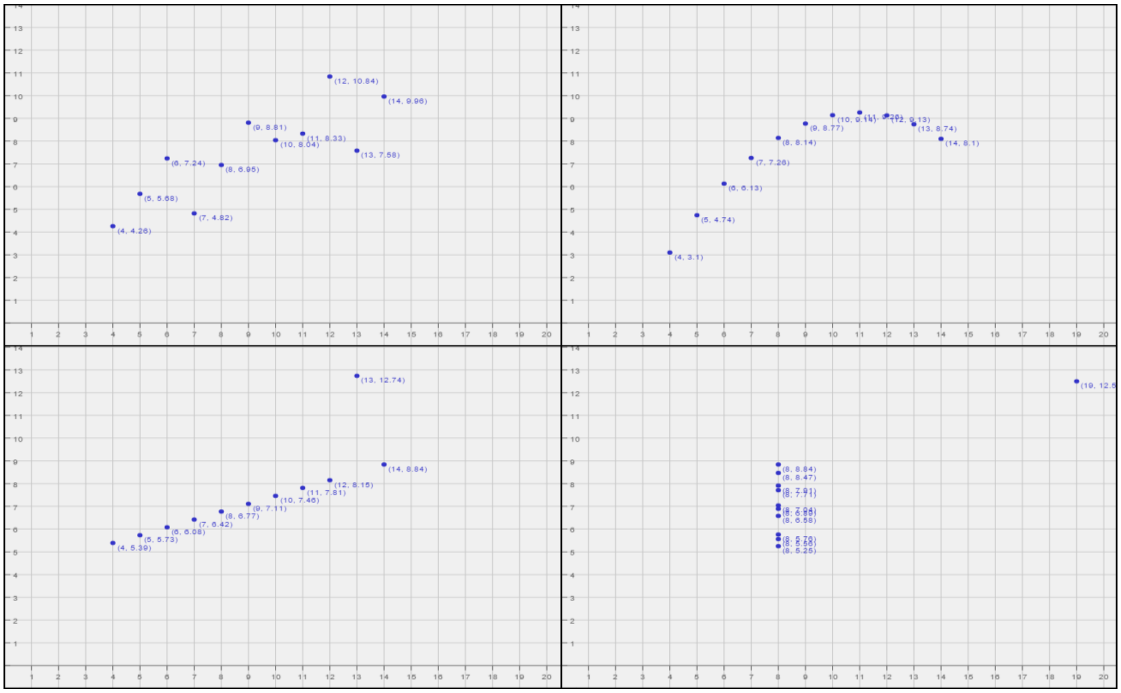
- **θ2:** coefficient of x

- Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

2. Explain the Anscombe's quartet in detail. (3 marks)

- Statistics are great for describing general trends and aspects of data, but statistics alone can't fully depict any data set. Francis Anscombe realized this in 1973 and created several data sets, all with several identical statistical properties, to illustrate it. These data sets, collectively known as “Anscombe's Quartet,” are shown in the picture below.

- All four of these data sets have the same variance in x, variance in y, mean of x, mean of y, and linear regression. But, as you can clearly tell, they are all quite different from one another. So, what does this mean to you as a statistician?
- Well, to start, Anscombe's Quartet is a great demonstration of the importance of graphing data to analyze it. Given simply variance values, means, and even linear regressions cannot accurately portray data in its native form. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed.
- Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets. Think about it: if the bottom two graphs didn't have that one point that strayed so far from all the other points, their statistical properties would no longer be identical to the two top graphs. In fact, their statistical properties would more accurately resemble the lines that the graphs seem to depict.
- *how to analyze your data*. For example, while all four data sets have the same linear regression, it is obvious that the top right graph really shouldn't be analyzed with a linear regression at all because it's a curvature. Conversely, the top left graph probably *should* be analyzed with a linear regression because it's a scatter plot that moves in a roughly linear manner. These observations demonstrate the value in graphing your data before analyzing it.
- Anscombe's Quartet reminds us that graphing data prior to analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety.



3. What is Pearson’s R? (3 marks)

- The **Pearson correlation coefficient** (**PCC**), also referred to as **Pearson's *r***, the **Pearson product-moment correlation coefficient** (**PPMCC**) or the **bivariate correlation**, is a measure of the linear correlation between two variables *X* and *Y*. According to the Cauchy–Schwarz inequality it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. It is widely used in the sciences. It was developed by Karl Pearson.

- Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier *product-moment* in the name.

$$\rho = \frac{Cov_{x,y}}{\sigma_x \sigma_y}$$

where

$$Cov_{x,y} = \frac{\sum(x-\overline{x})(y-\overline{y})}{n-1}$$

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

- **Feature scaling** is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.
- Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.
- Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.
- The terms *normalization* and *standardization* are sometimes used interchangeably, but they usually refer to different things. *Normalization* usually means to scale a variable to have a values between 0 and 1, while *standardization* transforms data to have a mean of zero and a standard deviation of 1. This standardization is called a **z-score**, and data points can be standardized with the following formula:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Where:

- xi is a data point (x1, x2…xn).
- x̄ is the sample mean.
- s is the sample standard deviation

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

- The ***variance inflation factor*** *(VIF)* quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing *collinearity/multicollinearity*. Higher values signify that it is difficult to assess accurately the contribution of predictors to a model.

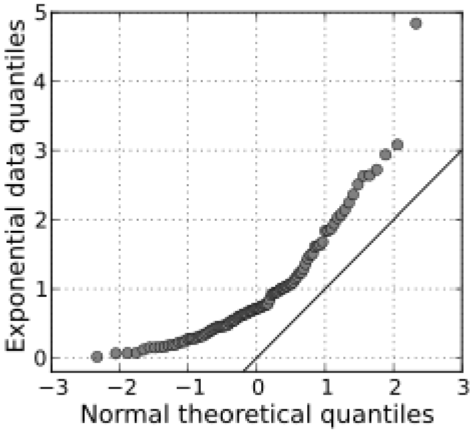
- **If there is perfect correlation or the corresponding variable may be expressed exactly by a linear combination of other variables then VIF = infinity.**

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

   - The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from population with a common distribution.

   - A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

   - A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.



The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

Understanding q-q Plot for Linear Regression Analysis

After running a regression analysis, we should check if the model works well for data. We can check if a model works well for data in many different ways. We pay great attention to regression results, such as slope coefficients, p-values, or R2 that tell us how well a model represents given data. That’s not the whole picture though. Residuals could show how poorly a model represents data. Residuals are leftover of the outcome variable after fitting a model (predictors) to data and they could reveal unexplained patterns in the data by the fitted model. Using this information, not only could we check if linear regression assumptions are met, but we can improve our model in an exploratory way. Diagnostic Plot like q-q plots helps to check whether model works for the data or not.