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)

   We could convert the categorical variables into dummies which follows (k-1) new dummies as they require a different technique so that training the model becomes simpler and more efficient.

   In our use case bike sharing, below are the few observations about categorical variables w.r.t dependent variable (cnt - total rental bikes)

   - season, yr, mnth, holiday, weekday, workingday and weathersit are the categorical variables available in the dataset
   - we see good number of bikes rented in fall and summer respectively as people tend to be more outdoors during that season, but it is less during winter and spring due to cold
   - we see more bikes are rented in 2019 compared to 2018 may be due more awareness on the rentals
   - Aligning to the seasons, we see bike rentals having high numbers between April and October and low between November to March
   - Assuming 0 as holiday and 1 not a holiday, we see people tend rent more during holidays
   - The days of a week did not really influence bike rentals, as they show similar trend across
   - Even working or non-working day does not much influence the bike rentals, as people tend to rent bikes for various purpose not just office commute
   - We see low bike rentals during light snow as it quite difficult to ride the bike during that time.

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

   The drop_first=True parameter in the pandas.get_dummies() function helps to reduce the number of dummy variables created and the correlations between them.

   As an instance if a variable has three unique values, you will create three respective dummy variables. This leads couple of problems

   Multicollinearity: A situation where 2 or more variables are highly correlated, this occurs when we create a dummy variable for each unique value of a variable

   Interpretability: having all dummy variables and not skipping the first one can make the interpretation of the regression coefficients less intuitive. When the first category is used as a reference, the coefficient estimates for the remaining dummy variables represent the change in the outcome variable relative to the reference category.

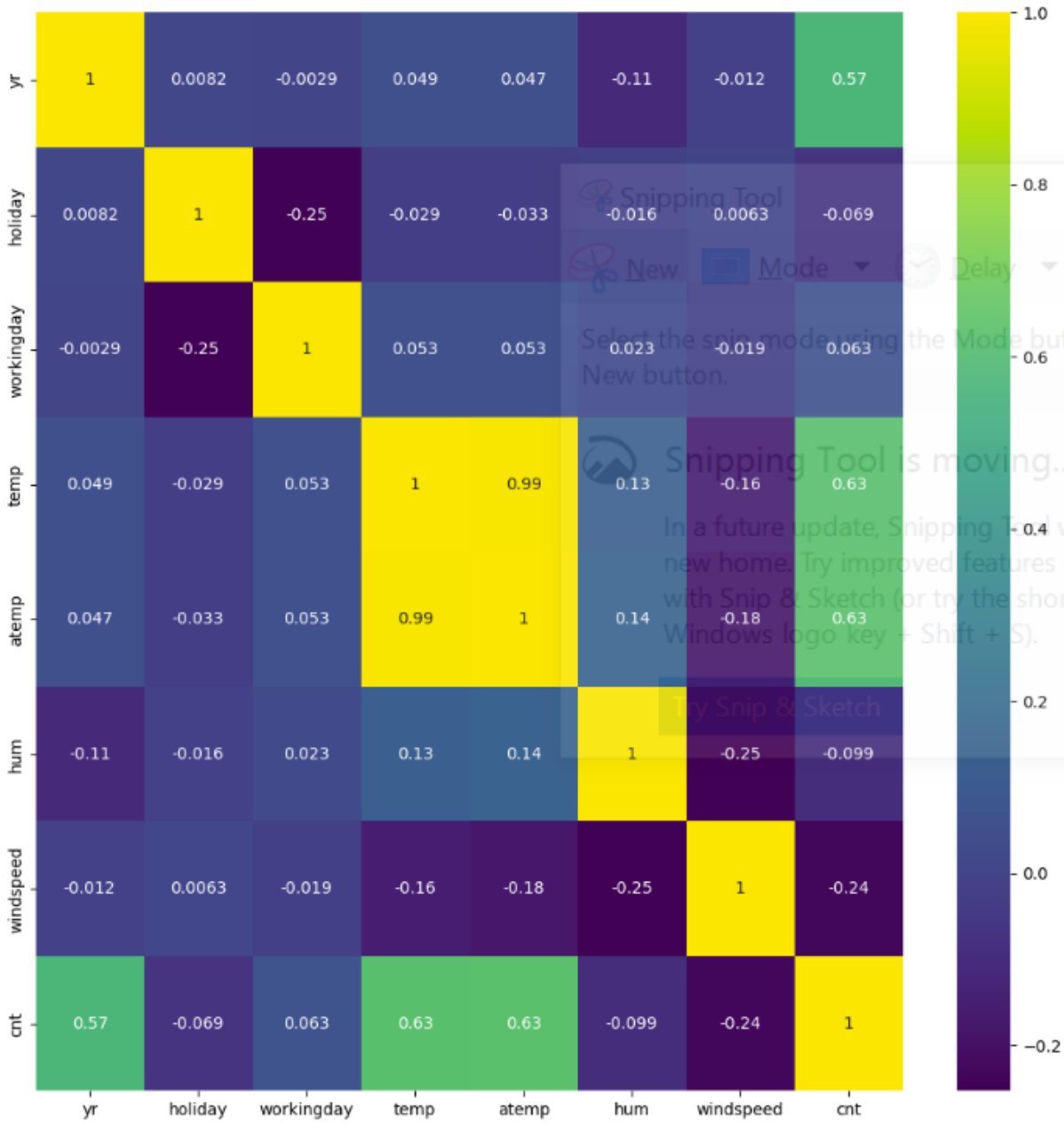
Also, it might cause overfitting as there too many variables for the model to generalize to new data

Owning to above, we leverage drop_first=True while dummy variable creation

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

   As cnt is the target variable, the highest correlation is atemp and next highest is temp.

   Refer to the screen on the correlation coefficient.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   The below assumptions are there to ensure our model is the best fit.

   Normality of residuals: One can find the normality of the residuals by plotting a histogram the residuals. it suggests that the assumption of normality is met when it follows a roughly normal distribution,

   Homoscedasticity: Homoscedasticity assumes that the residuals have a constant variance across all levels of the independent variables. One can find by plotting the residuals against the predicted values. If the spread of residuals is roughly constant across the range of predicted values, homoscedasticity is likely met. Otherwise, it violates the assumption.

   Linearity: A linear relationship between the independent variables and the dependent variable. One can find by plotting the actual values against the predicted values. If the points fall roughly along a straight line, it suggests that the linearity assumption holds.

   Multicollinearity: Linear regression assumes that the independent variables are not highly correlated with each other. To find multicollinearity we use the correlation matrix or VIF method

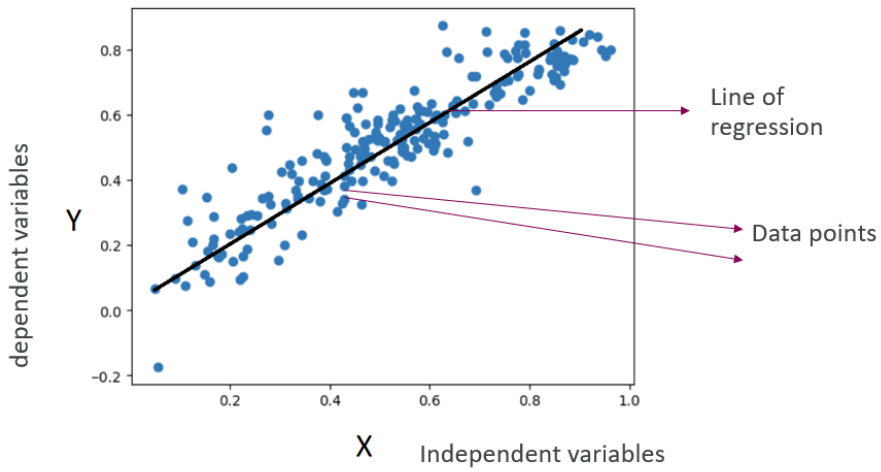
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   - feeling temperature: As feeling temp increase there is increase in bike rentals 0.4117 times and it is positive correlation.
   - Light Snow: As more the snow there is decrease in the bike rentals by 0.2912 and it is a negative correlation
   - Year: There is a increase in rentals for 2019 compared to 2018 …Year over year increase by 0. 2357 time and it is a positive correlation.

General Subjective Questions

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

   Linear regression is a supervised machine learning algorithm that is used to predict a continuous value (a real number) based on a set of independent variables. The independent variables are also known as the features or predictors, and the dependent variable is the target variable that you want to predict.

   The linear regression algorithm finds a linear relationship between the independent variables and the dependent variable. A linear relationship is a straight-line relationship, where the dependent variable changes linearly w.r.t the independent variable change.



   The equation for a linear regression line is: y = b0 + b1x

   y = dependent variable

   b0 = y-intercept

   b1 = slope of the line

   x = independent variable

   The y-intercept is the value of y when x is 0. The slope of the line tells you how much y changes for every unit change in x.

Here are some of the advantages of using linear regression:

- It is a simple and easy-to-understand algorithm
- It can be used with a variety of data types
- It is relatively efficient to compute
- It can be used to solve a variety of problems

Here are some of the disadvantages of using linear regression:

- It assumes that the relationship between the independent variables and the dependent variable is linear
- It can be sensitive to outliers
- It can be difficult to interpret the results of a linear regression model

Overall, linear regression is a powerful and versatile tool that can be used to make predictions about a variety of phenomena. It is a good choice for problems where you have a continuous dependent variable, and you believe that the relationship between the independent variables and the dependent variable is linear.

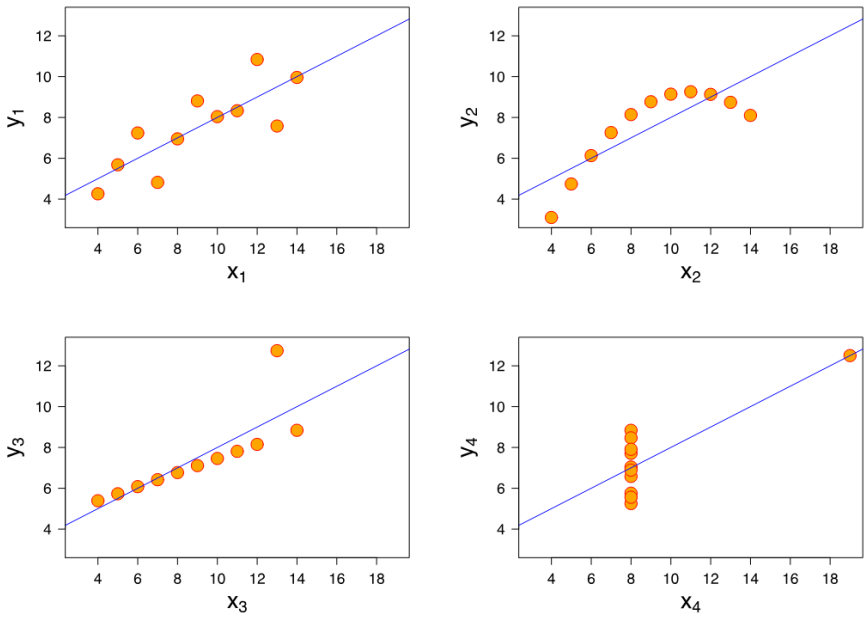
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a set of four data sets that have identical means, standard deviations, correlation coefficients, and regression lines, but appear very different when plotted. This demonstrates the importance of looking at data visually, as well as using statistical tests, in order to understand the relationships between variables.

The four data sets in Anscombe's quartet are:

Set 1: linear dataset with a positive slope.
Set 2: quadratic dataset with a positive slope.
Set 3: dataset with a strong positive correlation, but is not linear.
Set 4: dataset with a weak positive correlation and a large outlier.

When plotted, the four data sets look very different:



As you can see, Set 1 is a clear linear relationship, Set 2 is a quadratic relationship, Set 3 is a curved relationship, and Set 4 is a linear relationship with an outlier. This demonstrates that even though the four data sets have identical summary statistics, they can be very different when plotted.

Anscombe's quartet used for understanding the importance of looking at data visually. It can help you to identify relationships between variables that would not be obvious from summary statistics alone. It can also help you to identify outliers and other data points that may be affecting the results of your analysis.

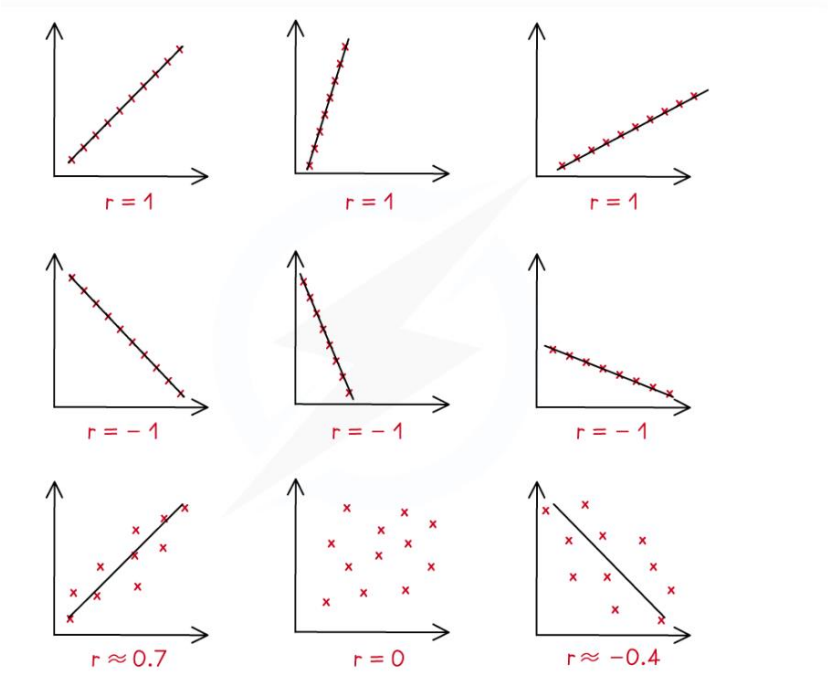
3. What is Pearson’s R? (3 marks)

Pearson's correlation coefficient (R) is a statistical measure that quantifies the strength and direction of the linear relationship between two variables.
Pearson's R is a value that ranges between -1 and +1. A correlation coefficient of +1 indicates a perfect positive linear relationship, where the variables increase or decrease together in a linear fashion. A correlation coefficient of -1 indicates a perfect negative linear relationship, where one variable increase while the other decreases in a linear fashion. A correlation coefficient of 0 indicates no linear relationship between the variables.

The formula for Pearson's correlation coefficient is as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r	=	correlation coefficient
xi	=	values of the x-variable in a sample
x bar	=	mean of the values of the x-variable
yi	=	values of the y-variable in a sample
y bar	=	mean of the values of the y-variable



- If the value of "r" is equal to zero, then it indicates that there is no correlation between the variables
- If the value of "r" is equal to 1 or -1, then it shows that there is a perfect positive or perfect negative correlation between the two variables.
- If the value of r is closer to 1 or -1, then it shows that there is a strong correlation between the given variables

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

   Scaling is a technique used in ML to bring the numerical values of different features or variables onto a similar scale. It involves transforming the original values of the features to a standardized range, making them comparable and easier to work with during modeling.

   Scaling is performed for several reasons like:

   1. Normalization: helps to normalize the data, ensuring that all features have a similar range. Which is important because many ML algorithms assume that the features are on a similar scale, and features with larger values can dominate the learning process.

   2. Distance-based algorithms: Scaling is particularly crucial when using distance-based algorithms, such as k-nearest neighbors (KNN) or support vector machines (SVM). These algorithms compute distances between data points, and if the features have different scales, the distances can be biased towards features with larger values.

   3. Gradient-based optimization: Scaling is essential for gradient-based optimization algorithms, such as gradient descent, which are commonly used in training neural networks. Scaling ensures that the optimization process is not hindered by features with large value ranges, as it can lead to slow convergence or the algorithm getting stuck in local optima.

   There are two common methods of scaling:

   1) Normalized Scaling:

      Normalized scaling also called Min-Max Scaling transforms the data to a specific range, typically between 0 and 1. It works by subtracting the minimum value of the feature and dividing it by the range (maximum value minus the minimum value).

      Formula:

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

      This scaling method preserves the shape of the distribution and is suitable when the distribution of the feature is not Gaussian.

2) Standardized Scaling:

Standardized scaling also called Z-score Normalization transforms the data to have a mean of 0 and a standard deviation of 1. It works by subtracting the mean of the feature and dividing it by the standard deviation. The formula for standardized scaling is:

Formula:

X_scaled = (X - mean) / standard_deviation

This scaling method assumes that the feature follows a Gaussian distribution. Standardized scaling helps to center the data around zero and brings it to a standard scale, where most values fall within the range of -3 to +3.

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

(3 marks)

A VIF of infinity indicates that there is perfect multicollinearity between two independent variables. This means that the two variables are perfectly correlated with each other, and there is no information gained by including both variables in the model.

Reasons for VIF being infinite

- The two independent variables are perfectly correlated with each other. This can happen if the two variables are measuring the same thing, or if they are highly correlated with a third variable that is not included in the model.
- The model has too many independent variables. When there are too many independent variables, it is possible for any two variables to be perfectly correlated with each other. This can happen even if the independent variables are not perfectly correlated with each other when considered in isolation.

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

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess the distributional similarity between two datasets. In the context of linear regression, a Q-Q plot is often employed to evaluate the assumption of normality for the residuals of the regression model.

Importance and use:

- A Q-Q plot in linear regression lies in its ability to detect deviations from the normality assumption, which is crucial for the validity of statistical inference procedures and parameter estimation. If the residuals are not normally distributed, it can impact the accuracy and reliability of the regression model. Violations of the normality assumption may lead to biased coefficient estimates, incorrect p-values, inaccurate confidence intervals, and misleading hypothesis testing results.

The normal q-q plot looks like below