# 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:

I found that there are 6 categorical variables in the dataset.

1. Seasons: the rentals are consistently high, with Fall obtaining a slight edge over others.
2. Year: Irrespective of the year, 2018 or 2019, the rentals seem consistent
3. Month: All the months have a good coverage of rentals with February falling short.
4. Holiday: There is a dramatic drop on holidays.
5. Day\_type: Weekdays show drastic increase in rentals while it is half short on weekends.
6. Weathersit : Users are seem worthwhile to choose bike rental on a pleasant weather, few short during a Gentle weather and seem abnormal on a Rough weather.

A group of different colored bars

Description automatically generated with medium confidence

2. Why is it important to use drop\_first=True during dummy variable creation?

Ans:

The drop\_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlation created among dummy variables. If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables. Consider a Categorical column weathersit (one of the categories)A screenshot of a computer program

Description automatically generated

As you can see in the screenshot, Fall has value 1 and sprint, summer and Winter have 0s.

Let's say we if drop the column 'Fall' then let's see how it can be explained by remaining three:

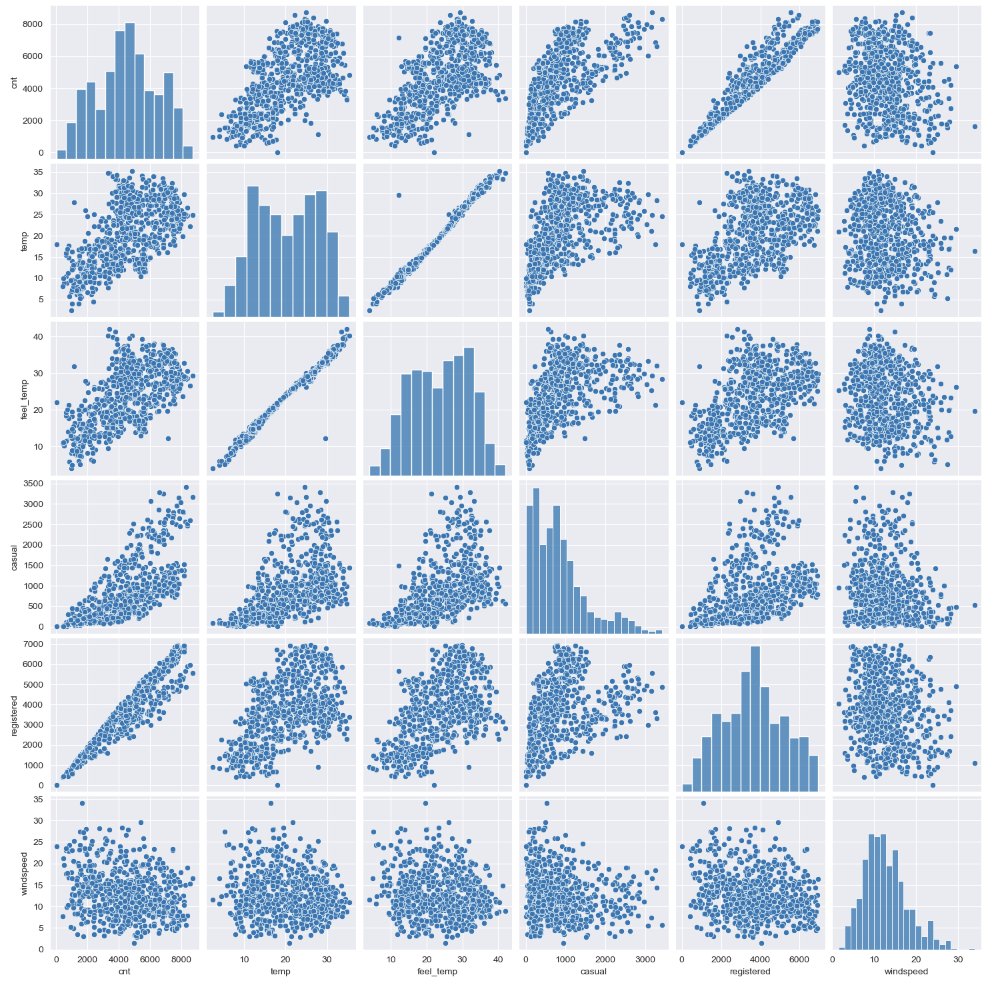
* 100 will corresponds to spring
* 010 will corresponds to summer
* 000 will corresponds to fall
* 001 will corresponds to winter

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

Ans:

There is a robust correlation between temp and Feel\_temp.

There is also a strong correlation between Casual and registered users.



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

Ans:

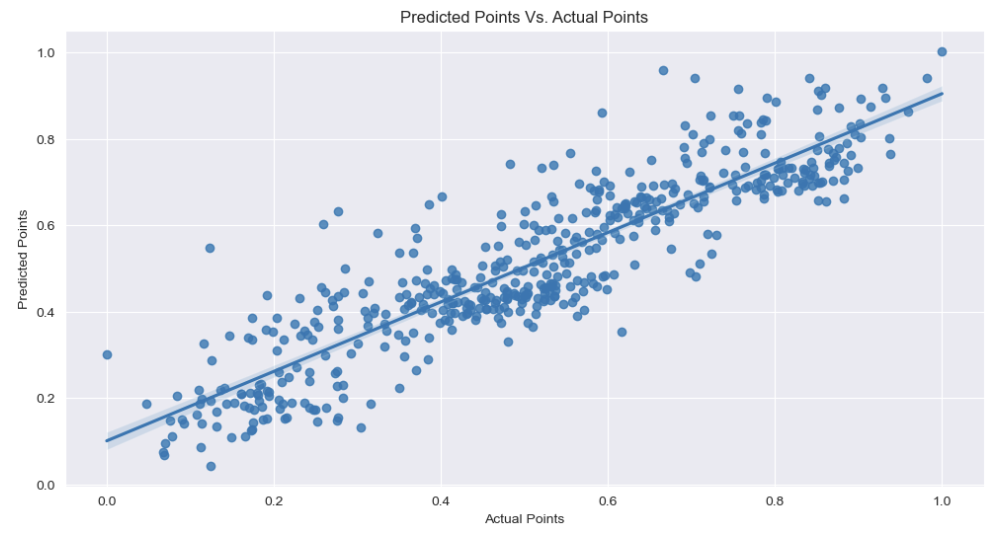
The assumptions of Linear Regression can be validated after building the model on the training set

1. The Error distribution follows the expected pattern and is 0 with high density with a normal tail.

A graph with a blue line

Description automatically generated

1. The Actual Vs Predicted values are consistent and close to the slope.



1. There is a constant deviation from the zero line, and there are no visible patterns in the error terms.

A graph showing a number of blue dots

Description automatically generated

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

* Casual Customers (Casual): Casual or not registered users are a strong group that the company needs to focus on. The company should also come up with strategy to register them.
* Year (year): The year has a positive impact on demand. Over time, there has been an increasing trend in bike rentals.
* September (Month): The month of September is associated with increased demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task, which means it predicts a continuous output variable (y) based on one or more input variables (x). It is mostly used for finding out the linear relationship between variables and forecasting. The basic idea of linear regression is to find a line that best fits the data points, such that the distance between the line and the data points is minimized. The line can be represented by an equation of the form:

y = θ0 + θ1x where θ0 is the intercept (the value of y when x is zero) and θ1 is the slope (the change in y for a unit change in x). These are called the parameters or coefficients of the linear model.

To find the best values of θ0 and θ1, we need to define a cost function that measures how well the line fits the data. A common choice is the mean squared error (MSE), which is the average of the squared differences between the actual y values and the predicted y values:

MSE = (1/n) \* ∑(y - y’)^2

where n is the number of data points, y is the actual value, and y’ is the predicted value.

The goal is to minimize the MSE by adjusting θ0 and θ1. There are different methods to do this, such as gradient descent, normal equation, or using libraries like scikit-learn.

Linear regression can also be extended to multiple input variables (x1, x2, …, xn), in which case the equation becomes:

y = θ0 + θ1x1 + θ2x2 + … + θnxn

Limitations are: it assumes a linear relationship between the input variables and the output variable, which may not always be the case. Another limitation is that it may be sensitive to outliers or multicollinearity.

2. Explain the Anscombe’s quartet in detail.

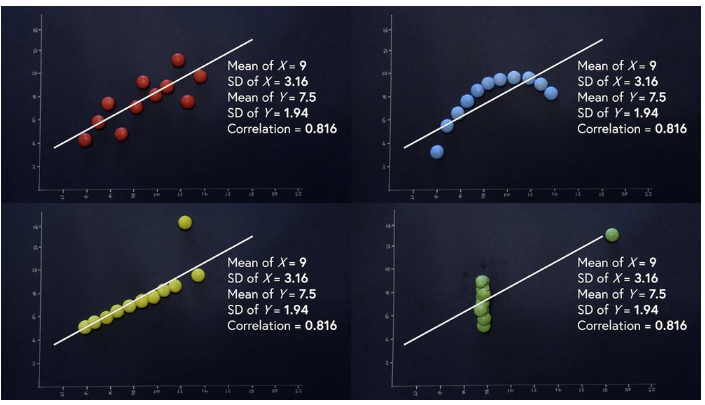
Anscombe’s Quartet shows how four entirely different data sets can be reduced down to the same summary metrics.

Here are the data sets from Anscombe’s Quartet – both as raw data, and plotted on a chart:

A table with numbers and symbols

Description automatically generated

* Mean of *X* = 9
* Standard deviation of *X* = 3.16
* Mean of *Y* = 7.5
* Standard deviation of *Y* = 1.94
* Correlation between *X & Y*= 0.816
* The linear regression (the line of best fit) is also the same



3. What is Pearson’s R?

Pearson’s correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Questions Answered:

Do test scores and hours spent studying have a statistically significant relationship?

Is there a statistical association between IQ scores and depression?

Assumptions:

**Independent of case**: Cases should be independent to each other.

**Linear relationship:** Two variables should be linearly related to each other. This can be assessed with a scatterplot: plot the value of variables on a scatter diagram, and check if the plot yields a relatively straight line.

**Homoscedasticity:** the residuals scatterplot should be roughly rectangular-shaped.

Properties:

**Limit:** Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists..

**Pure number:** It is independent of the unit of measurement. For example, if one variable’s unit of measurement is in inches and the second variable is in quintals, even then, Pearson’s correlation coefficient value does not change.

**Symmetric:** Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Degree of correlation:

**Perfect:** If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).

**High degree:** If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.

**Moderate degree:** If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.

**Low degree:** When the value lies below + .29, then it is said to be a small correlation.

**No correlation:** When the value is zero.

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

Scaling

Feature scaling is one of the most important data preprocessing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled. Tree-based algorithms are insensitive to the scale of the features. Also, feature scaling helps machine learning, and deep learning algorithms train and converge faster. There are some feature scaling techniques such as Normalization and Standardization that are the most popular and at the same time, the most confusing ones.

**Normalization or Min-Max Scaling** 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. Usually, we would scale age and not incomes because only a few people have high incomes but the age is close to uniform

**Standardization or Z-Score Normalization** is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X\_new = (X - mean)/Std

standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. 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.

**Difference between Normalization and Standardization**

|  |  |
| --- | --- |
| **Normalization** | **Standardization** |
| Minimum and maximum value of features are used for scaling | Mean and standard deviation is used for scaling. |
| It is used when features are of different scales. | It is used when we want to ensure zero mean and unit standard deviation. |
| Scales values between [0, 1] or [-1, 1]. | It is not bounded to a certain range. |
| It is really affected by outliers. | It is much less affected by outliers. |
| Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |
| This transformation squishes the n-dimensional data into an n-dimensional unit hypercube. | It translates the data to the mean vector of original data to the origin and squishes or expands. |
| It is useful when we don’t know about the distribution | It is useful when the feature distribution is Normal or Gaussian. |
| It is a often called as Scaling Normalization | It is a often called as Z-Score Normalization. |

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

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

The formula for VIF is:

VIF*i*=1 /(1-*Ri*​^2)

**where:**​ *Ri*​^2 is Unadjusted coefficient of determination forregressing the ith independent variable on theremaining ones​  
In general terms,

VIF equal to 1 = variables are not correlated

VIF between 1 and 5 = variables are moderately correlated

VIF greater than 5 = variables are highly correlated

VIF are infinity when somevariables can create perfect multiple regressions on other variables.

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

A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a normal QQ plot when both sets of quantiles truly come from normal distributions.

1. The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
2. The importance of a Q-Q plot in linear regression lies in its ability to assess the normality assumption of residuals visually and quantitatively.
3. If the Q-Q plot shows a straight line, it provides evidence that the residuals are normally distributed, which is one of the key assumptions of linear regression.
4. On the other hand, if the Q-Q plot shows significant deviations from a straight line, it suggests that the normality assumption may not hold, and you may need to consider transformations or other methods to address outliers in your data before drawing conclusions from your regression analysis

A graph of a normal q-q plot

Description automatically generated