**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:**

Analysis done on categorical columns using the pairplot, boxplot, histplot and reg plot. Below are the few points we can infer about their effect on the dependent variables:

- Temperature has a significant impact on bike rentals, it's highly correlated with the count.
- Fall season has the highest demand for rental bikes.
- Demand for 2019 has grown from previous year, which shows good progress in terms of business.
- Demand is continuously growing each month till June. September has the highest demand and then it started decreasing as we approached the end of year.
- When it's a holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- Booking seemed to be almost equal either on working day or non-working day.
- The clear weathersit has the highest demand.
- Hum values are more scattered around. Demand decreases with increase in humidity.
- Demand is decreasing with increase in wind speed.
- Demands are attracted more in clear weather followed by misty weather.

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

**(2 mark)**

**Answer:**

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. It reduces the correlations created among dummy variables.
drop_first = False use to create one dummy variable for every level of the input categorical variable.
And if you set drop_first = True , then **it will drop the first category**. So if you have K categories, it will only produce K – 1 dummy variables.

**For example:**

Let's say we have 3 types of values in the Categorical column and we want to create a dummy variable for that column. If one variable is not furnished and second is semi_furnished, then third is obviously unfurnished. So we do not need a 3rd variable to identify the unfurnished.

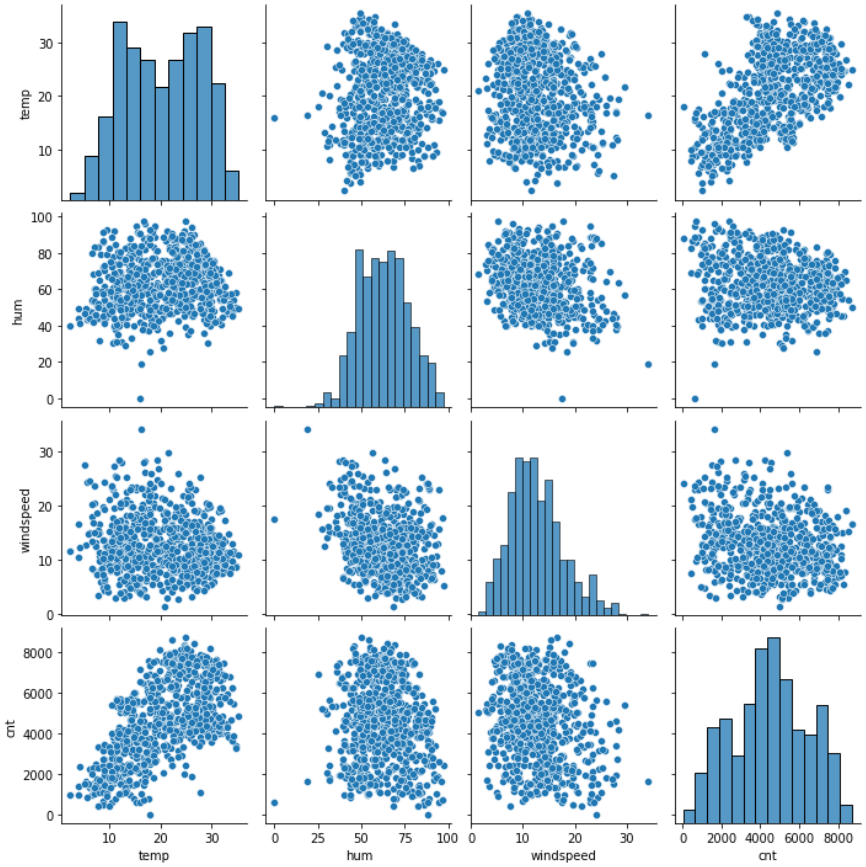
| Value | Indicator Variable | |
|---|---|---|
| Furnishing Status | furnished | semi-furnished |
| furnished | 1 | 0 |
| semi-furnished | 0 | 1 |
| unfurnished | 0 | 0 |

I have removed the unfurnished as per example above. Still we can tell what row is unfurnished looking at the data.

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

**Answer:**

In the pair-plot among numerical variables 'temp' has the highest correlation with the target variable.

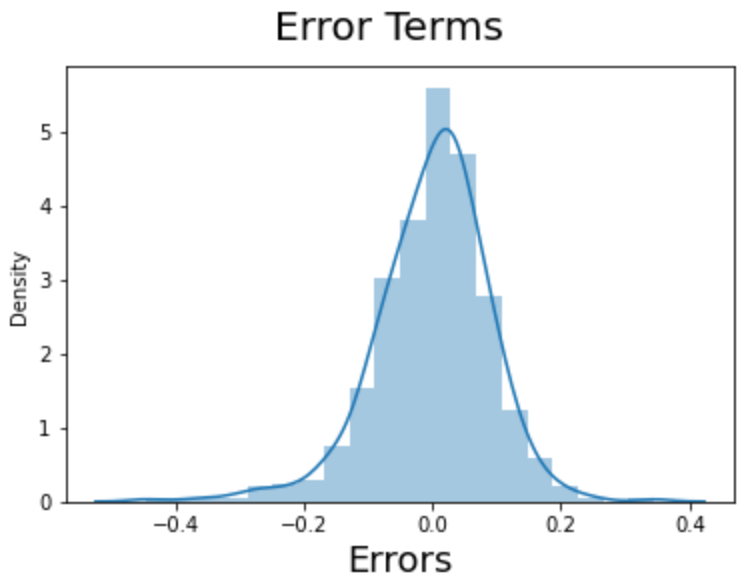



**4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

**Answer:**

I have validated the assumption of Linear Regression Model based on **normality of error terms.**
A residual plot is a type of plot that displays the fitted values against the residual values for a regression model.

Based on the histogram, we concluded that error terms are following a normal distribution. Normality of error terms should be normally distributed with mean zero.




**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:**

The top 3 features contributing significantly towards explaining the demand of the shared bikes are –

- **Temp** - Coefficient value of '`0.4387`' indicated that a unit increase in temp variable, means a temperature has a significant impact on bike rentals.
- **Yr** - Coefficient value of '`0.2345`' indicated that a unit increase in yr variable, that means year wise rental numbers are increasing.
- **Sept** - A coefficient value of '`0.0687`' indicated that a unit increase in Sep variable increases the bike hire numbers by `0.0687` units.

**General Subjective Questions**

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

**Answer:**

Linear regression is one of the oldest and most popular algorithms. With roots in the statistics world, the algorithm is used for solving regression problems. This means that the final output of the model is a numeric value. The algorithm maps a linear relationship between the input features(X) and the output (y). Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are multiple input variables, literature from statistics often refers to the method as **multiple linear regression**.

Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called **Ordinary Least Squares**.

**When and why do you use Regression?**

Regression is performed when the dependent variable is of continuous data type, and Predictors or independent variables could be of any data type like continuous, nominal/categorical, etc. The regression method tries to find the best fit line, which shows the relationship between the dependent variable and predictors with the least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

To find a linear relationship between independent and dependent variables by using a linear equation on the data.

The equation for a linear line is -

$$Y = mx + c$$

Where **m** is slope and **c** is the intercept.

In Linear Regression, we are actually trying to predict the best m and c values for dependent variable Y and independent variable x. We fit as many lines and take the best line that gives the least possible error. We use the corresponding m and c values to predict the y value.

The same concept can be used in multiple Linear Regression where we have multiple independent variables, x1, x2, x3…xn.

Now the equation changes to -

**Y = M1X1 + M2X2 +.....MnXn + C**

The above equation is not a line but a plane of multi-dimensions.

Linear Regression is the basic form of regression analysis.

There are four assumptions associated with a linear regression model:

1. **Linearity**: The relationship between independent variables and the mean of the dependent variable is linear.
2. **Homoscedasticity**: The variance of residuals should be equal.
3. **Independence**: Observations are independent of each other.
4. **Normality**: The dependent variable is normally distributed for any fixed value of an independent variable.

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

**Answer:**

**Anscombe's quartet** was 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**.**

Anscombe's Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.

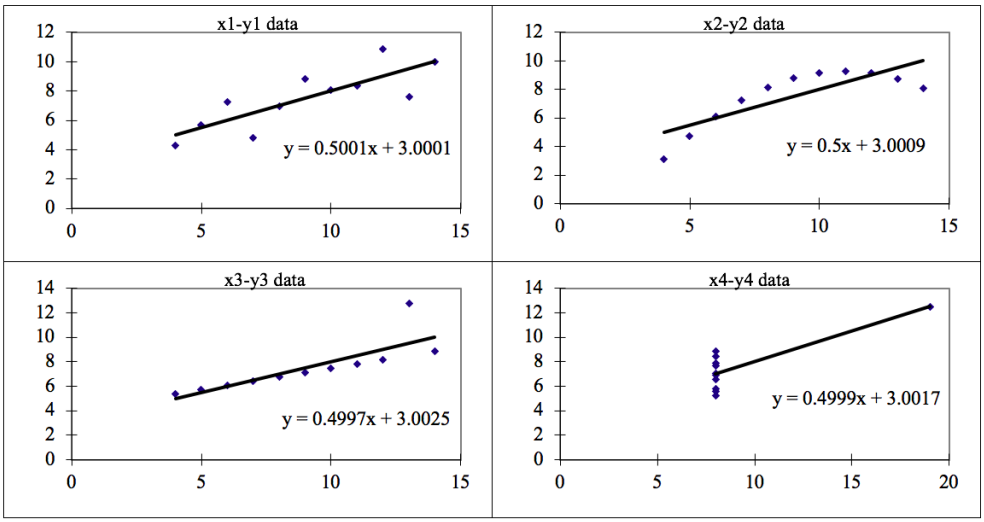
There are four data set plots which have nearly the s**ame statistical observations**, which provide the same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets. These four plots can be defined as follows:

| Anscombe's Data | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Observation | x1 | y1 | | x2 | y2 | | x3 | y3 | | x4 | y4 |
| 1 | 10 | 8.04 | | 10 | 9.14 | | 10 | 7.46 | | 8 | 6.58 |
| 2 | 8 | 6.95 | | 8 | 8.14 | | 8 | 6.77 | | 8 | 5.76 |
| 3 | 13 | 7.58 | | 13 | 8.74 | | 13 | 12.74 | | 8 | 7.71 |
| 4 | 9 | 8.81 | | 9 | 8.77 | | 9 | 7.11 | | 8 | 8.84 |
| 5 | 11 | 8.33 | | 11 | 9.26 | | 11 | 7.81 | | 8 | 8.47 |
| 6 | 14 | 9.96 | | 14 | 8.1 | | 14 | 8.84 | | 8 | 7.04 |
| 7 | 6 | 7.24 | | 6 | 6.13 | | 6 | 6.08 | | 8 | 5.25 |
| 8 | 4 | 4.26 | | 4 | 3.1 | | 4 | 5.39 | | 19 | 12.5 |
| 9 | 12 | 10.84 | | 12 | 9.13 | | 12 | 8.15 | | 8 | 5.56 |
| 10 | 7 | 4.82 | | 7 | 7.26 | | 7 | 6.42 | | 8 | 7.91 |
| 11 | 5 | 5.68 | | 5 | 4.74 | | 5 | 5.73 | | 8 | 6.89 |

The statistical information for all these four datasets are approximately similar and can be computed as follows:

| Anscombe's Data | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Observation | x1 | y1 | | x2 | y2 | | x3 | y3 | | x4 | y4 |
| 1 | 10 | 8.04 | | 10 | 9.14 | | 10 | 7.46 | | 8 | 6.58 |
| 2 | 8 | 6.95 | | 8 | 8.14 | | 8 | 6.77 | | 8 | 5.76 |
| 3 | 13 | 7.58 | | 13 | 8.74 | | 13 | 12.74 | | 8 | 7.71 |
| 4 | 9 | 8.81 | | 9 | 8.77 | | 9 | 7.11 | | 8 | 8.84 |
| 5 | 11 | 8.33 | | 11 | 9.26 | | 11 | 7.81 | | 8 | 8.47 |
| 6 | 14 | 9.96 | | 14 | 8.1 | | 14 | 8.84 | | 8 | 7.04 |
| 7 | 6 | 7.24 | | 6 | 6.13 | | 6 | 6.08 | | 8 | 5.25 |
| 8 | 4 | 4.26 | | 4 | 3.1 | | 4 | 5.39 | | 19 | 12.5 |
| 9 | 12 | 10.84 | | 12 | 9.13 | | 12 | 8.15 | | 8 | 5.56 |
| 10 | 7 | 4.82 | | 7 | 7.26 | | 7 | 6.42 | | 8 | 7.91 |
| 11 | 5 | 5.68 | | 5 | 4.74 | | 5 | 5.73 | | 8 | 6.89 |
| | | | | Summary Statistics | | | | | | | |
| N | 11 | 11 | | 11 | 11 | | 11 | 11 | | 11 | 11 |
| mean | 9.00 | 7.50 | | 9.00 | 7.500909 | | 9.00 | 7.50 | | 9.00 | 7.50 |
| SD | 3.16 | 1.94 | | 3.16 | 1.94 | | 3.16 | 1.94 | | 3.16 | 1.94 |
| r | 0.82 | | | 0.82 | | | 0.82 | | | 0.82 | |

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:




The four datasets can be described as**:**

- **Dataset 1:** this fits the linear regression model pretty well.
- **Dataset 2:** this could not fit linear regression model on the data quite well as the data is non-linear.
- **Dataset 3:** shows the outliers involved in the dataset which cannot be handled by a linear regression model.
- **Dataset 4:** shows the outliers involved in the dataset which cannot be handled by a linear regression model.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

**3. What is Pearson's R?** **(3 marks)**

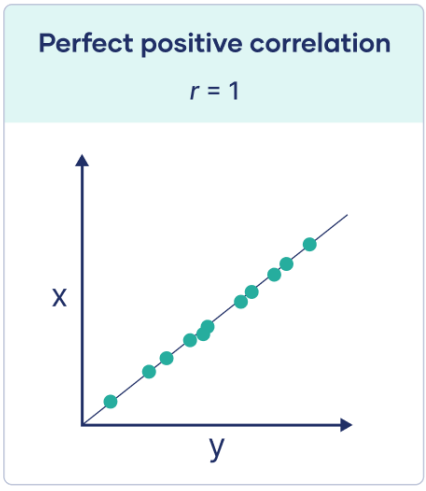
**Answer:**

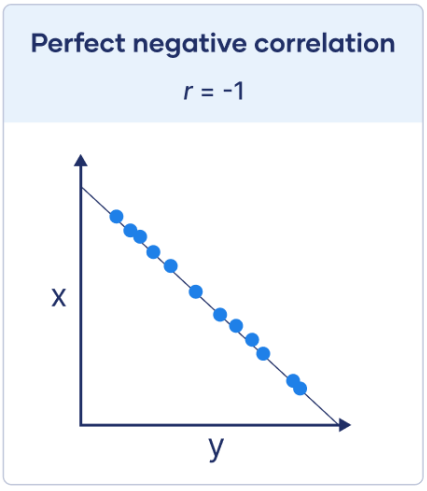
**Correlation coefficients** are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is **Pearson's**. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in **linear regression**.

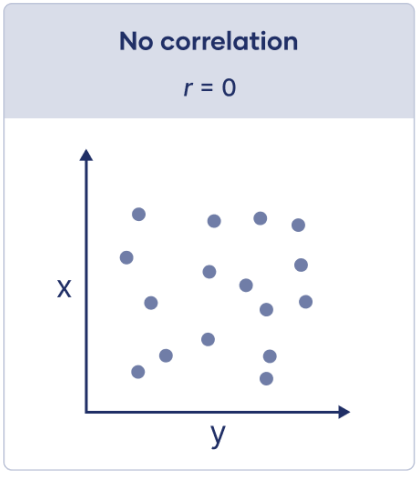
**Correlation is a statistic that measures the relationship between two variables in the finance and investment industries.** It shows the strength of the relationship between the two variables as well as the direction and is represented numerically by the correlation coefficient. The numerical values of the correlation coefficient lies between **-1.0 and +1.0.**

**When the value of the correlation coefficient is exactly 1.0, it is said to be a perfect positive correlation.** This situation means that when there is a change in one variable, either negative or positive, the second variable changes in lockstep, in the same direction.

**A perfect negative correlation means that two assets move in opposite directions, while a zero correlation implies no linear relationship**. We can determine the strength of the relationship between two variables by finding the absolute value of the correlation coefficient.



**Pearson correlation coefficient formula:**

Below is a formula for calculating the Pearson correlation coefficient (r):

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

Above mentioned formula can also be used in software such as R or Excel to calculate the Pearson correlation coefficient.

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

**Answer:**

**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.
**For example —** if multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Rs), and (1–2 Meters) respectively, feature scaling would help them all to be in the same range, for example - centered around 0 or in the range (0,1) depending on the scaling technique.

**Methods for Scaling:** The most common scaling methods are:

1. **Normalization -** Normalization, also known as **min-max scaling** or **min-max normalization**, is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

   Here, **max(x)** and **min(x)** are the maximum and the minimum values of the feature respectively.

2. **Standardization -** Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

**Why is scaling performed? -** Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

**Difference between normalized scaling and standardized scaling -**

| S. No. | Normalized scaling | Standardized scaling |
|---|---|---|
| **1** | Normalization is used when the data doesn't have Gaussian distribution. | Standardization is used on data having Gaussian distribution. |
| **2** | Normalization scales in a range of [0,1] or [-1,1]. | Standardization is not bounded by range. |
| **3** | Normalization is highly affected by outliers. | Standardization is slightly affected by outliers. |
| **4** | Normalization is considered when the algorithms do not make assumptions about the data distribution. | Standardization is used when algorithms make assumptions about the data distribution. |
| **5** | Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |
| **6** | Minimum and maximum value of features are used for scaling. | Mean and standard deviation is used for scaling. |

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

**Answer:**

**VIF -** A variance inflation factor(VIF) detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model; its presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

VIFs are calculated by taking a predictor, and regressing it against every other predictor in the model. This gives you the R-squared values, which can then be plugged into the VIF formula.

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

Where, 'i' refers to the ith variable.

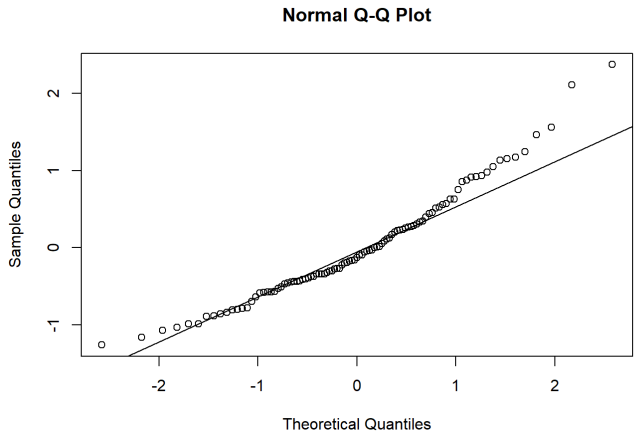
***If the R-squared value is equal to 1 then the denominator of the above formula becomes 0 and the overall value becomes infinite. It denotes perfect correlation in variables.***

***If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, 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)**

**Answer:**

**Q-Q plot:** Q-Q plots are also known as **Quantile-Quantile plots.** As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.
Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.
**For example,** if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.




The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q Q plot is called a normal quantile-quantile (QQ) plot. The points are not clustered on the 45 degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.

**Use of QQ plots:** It is very useful to determine -

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution - If the see the left side of the plot deviating from the line, it is left-skewed. When the right side of the plot deviates, it's right-skewed.

**Importance:** When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.