Assignment-based Subjective

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-1)The demand of bike is less in the month of spring when compared with other seasons**

**2)The demand for bikes increased in the year 2019 when compared with the year 2018.**

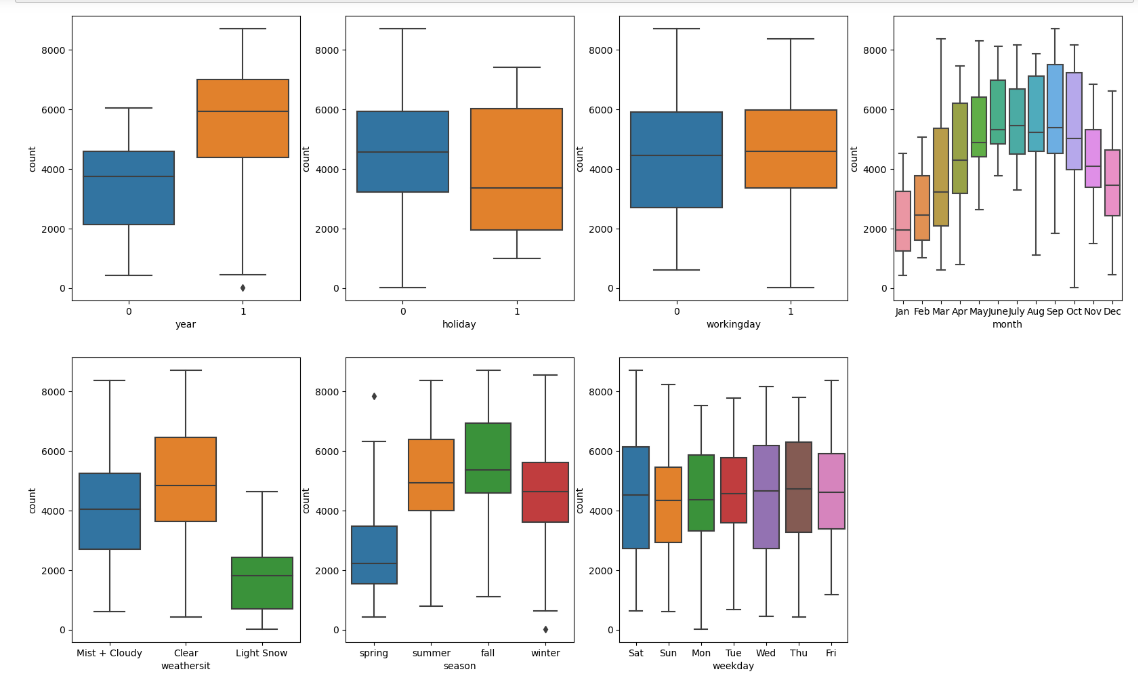
**3)Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.**

**4)Bike demand is less in holidays in comparison to not being holiday.**

**5)The demand for bikes is almost similar throughout the weekdays.**

**6)There is no significant change in bike demand with working days and non working days.**

**7)The bike demand is high when the weather is clear and Few clouds, however demand is less in case of Lightsnow and light rainfall.**

**8)Bike demand in the fall is the highest.**

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

**Ans- it is important in order to achieve k-1 dummy variables as it can be used to delete extra columns while creating dummy variables.**

**For Example: We have three variables: Furnished, Semi-furnished and unfurnished.We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, sowe don’t need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it**

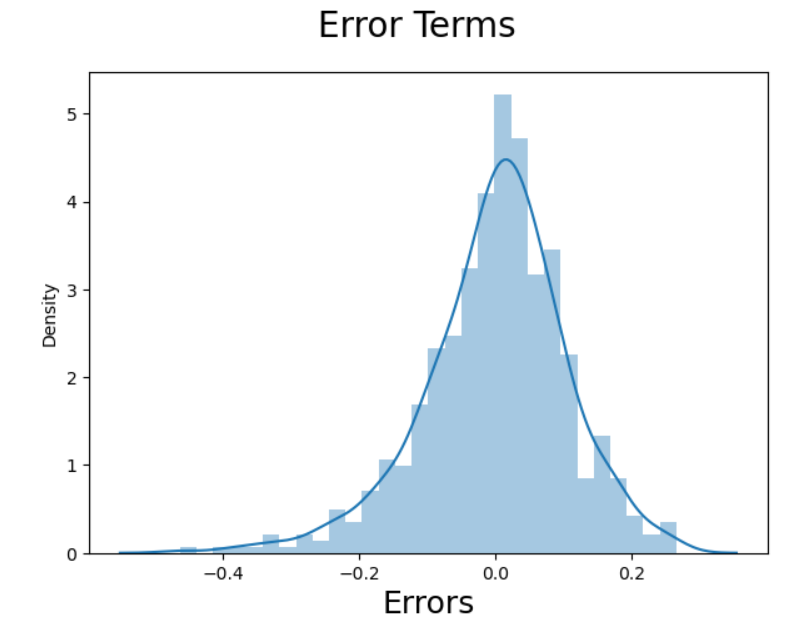
**It is also used to reduce the collinearity between dummy variables**

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 both have the same correlation with the target variable of 0.63 which is the highest among all numerical variables.**

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

**ans-According to this assumption there is a linear relationship between the features and target. Linear regression captures only linear relationships. This can be validated by using residual analysis(Plot the histogram of the error terms).**



**From the above histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.**

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-Top 3 features contributing significantly towards the demands of share bikes are:**

**1)weathersit\_Light\_Snow(negative correlation).**

**2)yr\_2019(Positive correlation).**

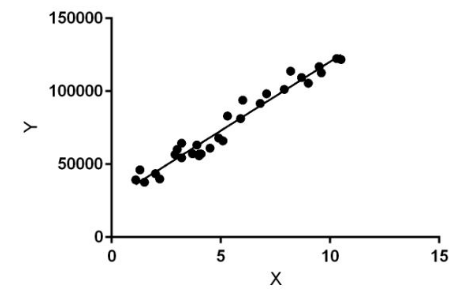
**3)temp(Positive correlation).**

General Subjective

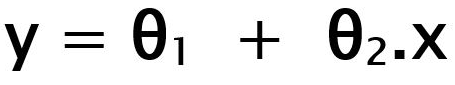
Questions

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

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model. Hypothesis function for Linear Regression :



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.

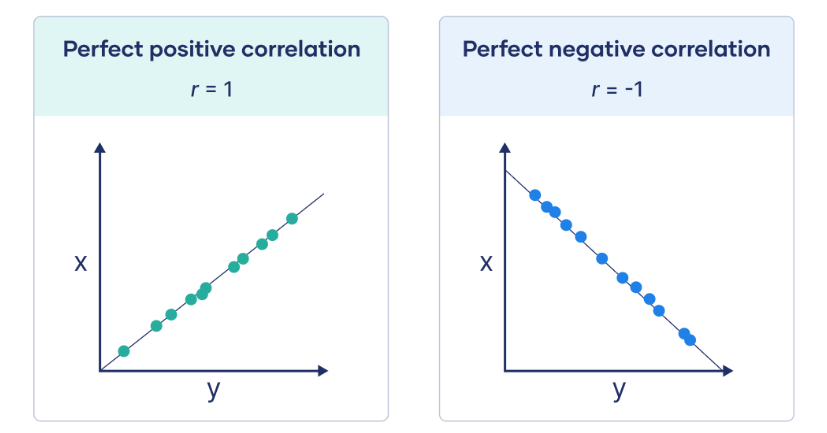
2. Explain the Anscombe’s quartet in detail. (3 marks)

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.

3. What is Pearson’s R? (3 marks)

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. When one variable changes, the other variable changes in the same direction.

| Pearson correlation coefficient (r) | Correlation type | Interpretation | Example |
| --- | --- | --- | --- |
| Between 0 and 1 | Positive correlation | When one variable changes, the other variable changes in the same direction. | Baby length & weight:  The longer the baby, the heavier their weight. |
| 0 | No correlation | There is no relationship between the variables. | Car price & width of windshield wipers:  The price of a car is not related to the width of its windshield wipers. |
| Between  0 and –1 | Negative correlation | When one variable changes, the other variable changes in the opposite direction. | Elevation & air pressure:  The higher the elevation, the lower the air pressure. |



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

It is a step of **data Pre-Processing** which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

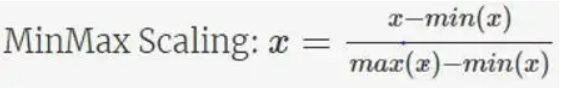
Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same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

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

**Normalization/Min-Max Scaling:**

It brings all of the data in the range of 0 and 1.

sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.



**Standardization Scaling:**

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

sklearn.preprocessing.scale helps to implement standardization in python.

**One disadvantage of normalization over standardization** is that it loses some information in the data, especially about outliers.

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

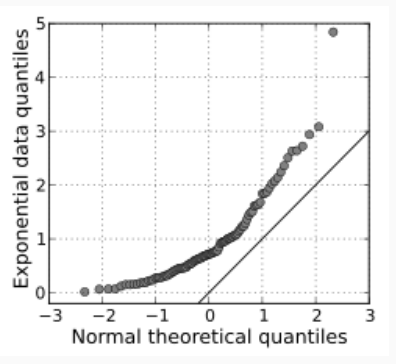
If there is perfect correlation, then **VIF = infinity**. This shows a **perfect correlation** between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.