

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?

Answer:

- The fall season appears to have attracted more bookings. The number of bookings has notably risen for each season in 2019 compared to 2018.
- Clear weather has led to more bookings, which is expected. Additionally, bookings increased in 2019 for all weather conditions compared to 2018.
- On non-holiday days, bookings tend to be lower, which makes sense as people might prefer to stay home and spend time with family during holidays.
- The number of bookings was nearly the same on both working days and non-working days. However, the count increased from 2018 to 2019.
- There were more bookings in 2019 compared to the previous year, as shown in earlier plots.
- Most bookings occurred in May, June, July, August, September, and October. The trend increased from the start of the year until mid-year, then started to decline towards the end of the year. Each month's bookings increased from 2018 to 2019.
- Wednesdays, Thursdays, Fridays, and Saturdays saw more bookings compared to other days (Sundays, Mondays, and Tuesdays) in 2019.
- The distinction between registered and casual users can only be made after a bike rental occurs, making both features functionally equivalent to the target variable. Consequently, these two features are redundant and can be removed.
- Registered Bike users are more than Casual and is strongly correlated with Target Variable
- Temp and atemp are very strongly correlated one can be dropped
- Humidity has not much Negative effect on the Target Variable.
- 'temp' and 'atemp' are both 63% correlated with the target variable 'cnt'
- 'temp' and 'atemp' are 99% correlated with each other. Both of the parameters cannot be used in the model due to multicollinearity issue. We will decide later which parameters to keep based on VIF and p-value w.r.t other variables.

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

Answer: The purpose of using dummy variables is to represent a categorical variable with 'n' levels by creating 'n-1' new columns, each indicating the presence or absence of a specific level using a zero or one. Setting `drop_first=True` ensures that the resultant columns match up to 'n-1' levels, reducing correlation among the dummy variables.

For example, if there are 3 levels, setting `drop_first=True` will drop the first column. When creating dummy variables for categorical data in regression analysis, it is crucial to use `drop_first=True` to avoid multicollinearity issues, known as the "dummy variable trap." The dummy variable trap occurs when the model has perfect multicollinearity, meaning one of the dummy variables can be perfectly predicted by the others.

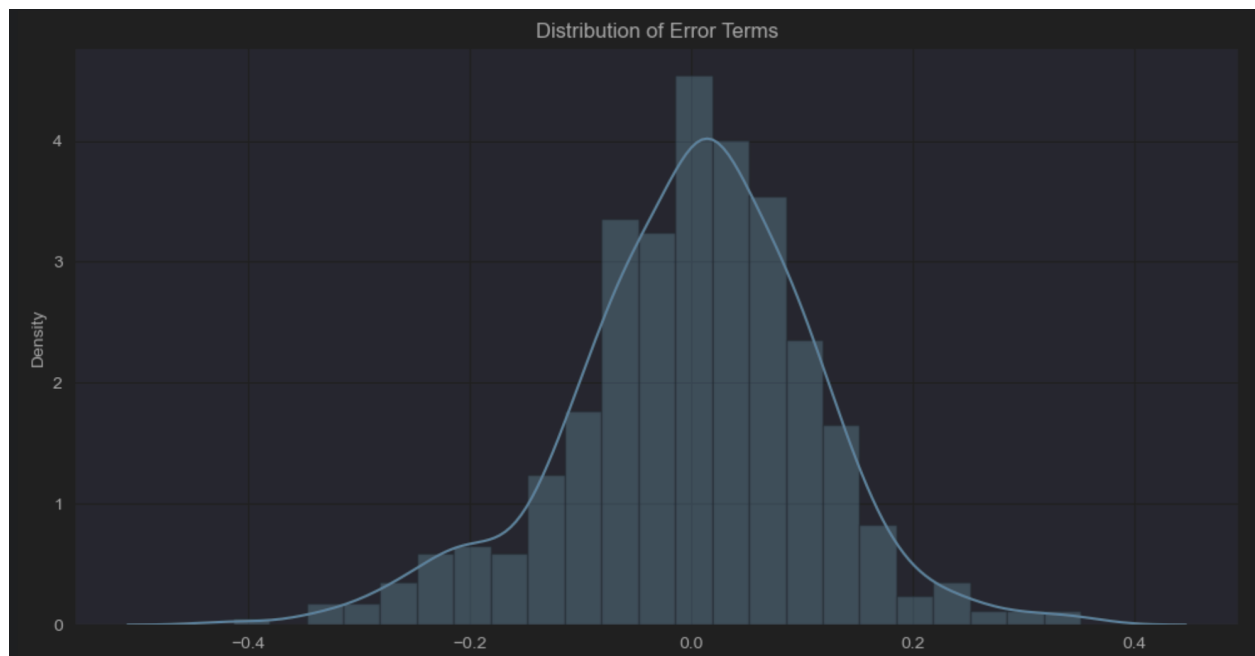
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: 'temp' variable 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?

Answer: Please find below details for assumptions related to the LR Model:

- *Normality of error terms*: Error terms should be normally distributed.



The error terms are fairly normally distributed with mean centered around zero

- *Multicollinearity check*: There should be insignificant multicollinearity among variables.

```
8 print(lr_model.summary())
9 print(get_variance_inflation_factor(X_train_rfe))
Executed at 2024.07.31 22:42:16 in 12ms
```

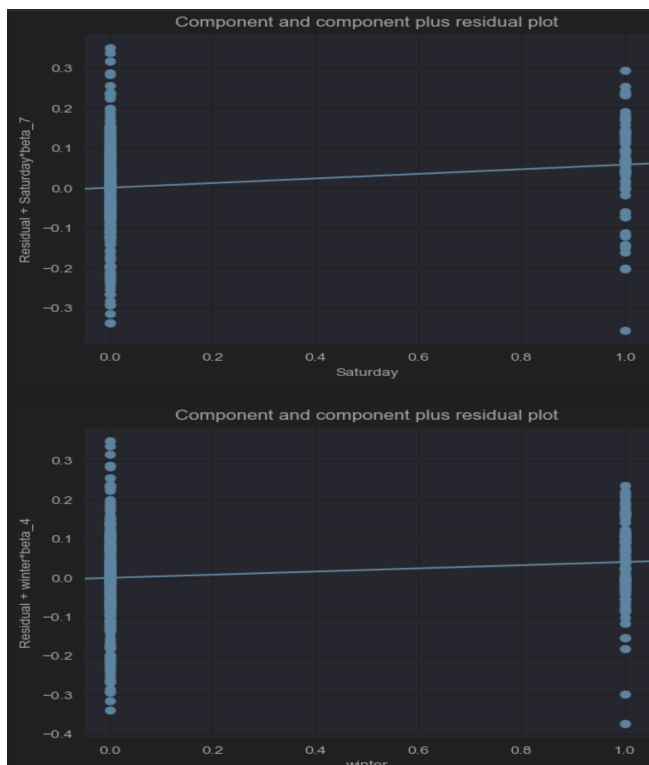
Kurtosis: 3.858 Cond. No. 9.09
=====

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

	Features	VIF
3	winter	2.28
1	workingday	2.19
0	yr	1.78
5	Nov	1.66
8	Mist	1.47
2	spring	1.41
4	Dec	1.33
6	Saturday	1.24
7	Light	1.08

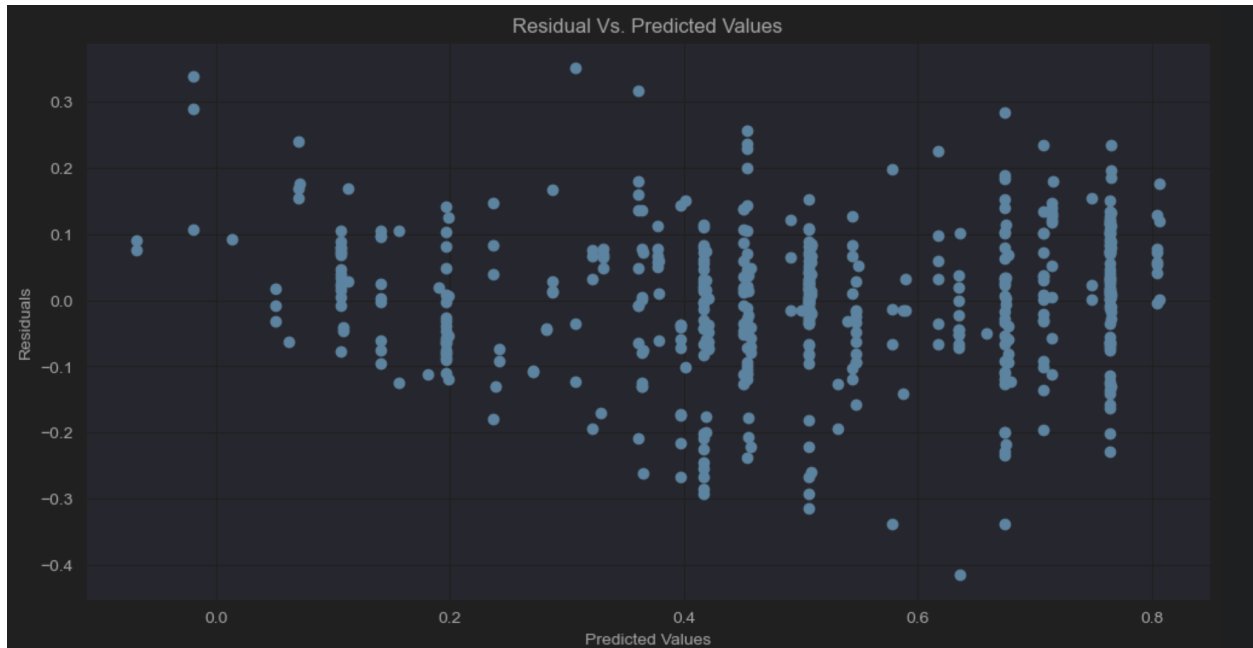
The details of VIF shows no multicollinearity.

- *Linear relationship validation*: Linearity should be visible among variables



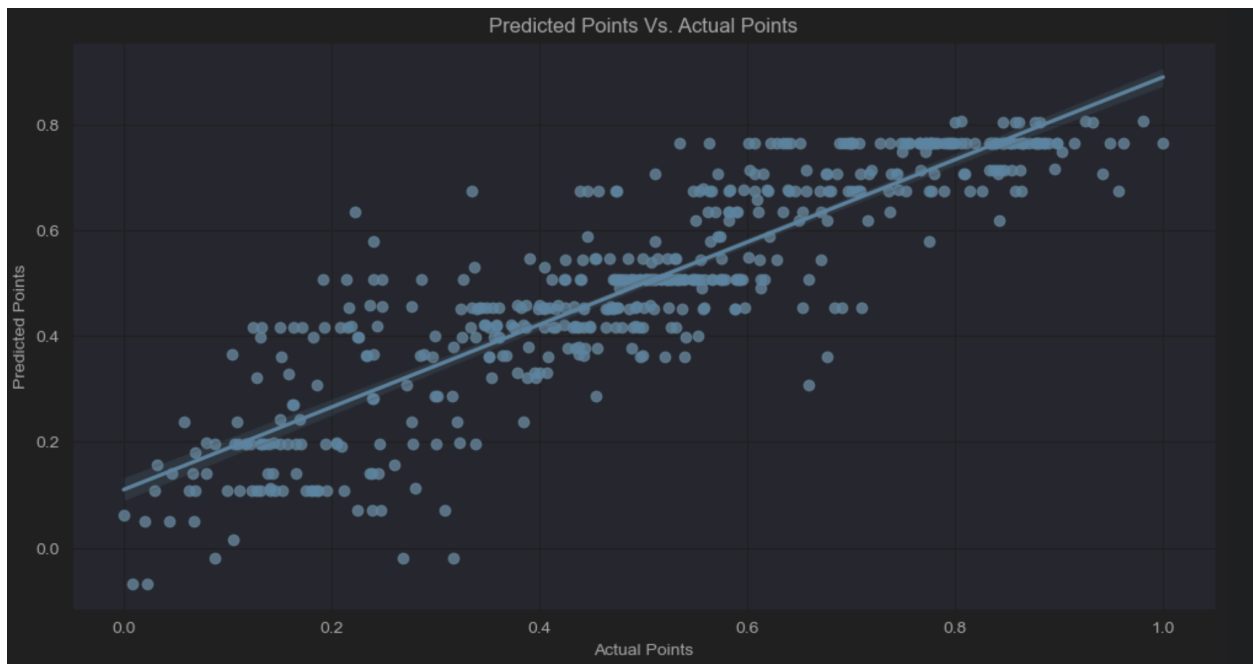
Linearity can be observed from the above plots.

- Homoscedasticity: There should be no visible pattern in residual values.



No visible pattern observed from the above plot for residuals.

- Independence of residuals:



No auto-correlation.

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

Answer:

1. Winter
2. sat
3. Workingday

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a form of predictive modeling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. If there is a single independent variable (x), such linear regression is called simple linear regression. And if there is more than one independent variable, such linear regression is called multiple linear regression.

Assumptions of Linear Regression:

1. Linearity: The relationship between the independent and dependent variables is linear.
2. Independence: The residuals (errors) are independent.
3. Homoscedasticity: The residuals have constant variance.
4. Normality: The residuals are normally distributed.
5. No Multicollinearity: The independent variables are not highly correlated. The method of least squares is used to find the best-fitting line for the observed data.
6. The estimated least squares regression equation has the minimum sum of squared errors, or deviations, between the fitted line and the observations.
7. The line of best fit is calculated by using the cost function – Least Sum of Squares of

Errors.

The objective of linear regression is to minimize the sum of the squared differences between the observed values and the predicted values (residuals). This method is known as Ordinary Least Squares (OLS).

Steps in Linear Regression:

1. Data Collection and Preparation: Gather and clean the data. Split the data into training and testing sets.
2. Exploratory Data Analysis (EDA): Visualize the data to understand relationships between variables. Check for linearity, correlation, and outliers.
3. Feature Selection and Engineering: Select relevant features that influence the dependent variable. Create new features if necessary.
4. Model Training: Use the training data to fit the linear regression model. Calculate the coefficients (β_0 , β_1 , ..., β_n).
5. Model Evaluation: Evaluate the model using the testing data. Common metrics: R-squared (R^2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE).
6. Prediction: Use the trained model to make predictions on new data.

2. Explain the Anscombe's quartet in detail.

Answer:

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. It was constructed to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

- 1st data set fits linear regression model as it seems to be linear relationship between X and y.
- 2nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.
- 3rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.
- 4th data set has a high leverage point means it produces a high correlation coeff.

Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning model.

Importance of Anscombe's Quartet:

Anscombe's quartet demonstrates several important points:

- Graphical Analysis: Summary statistics alone can be misleading. Visualizing data is crucial for identifying patterns, relationships, and anomalies that statistics might miss.

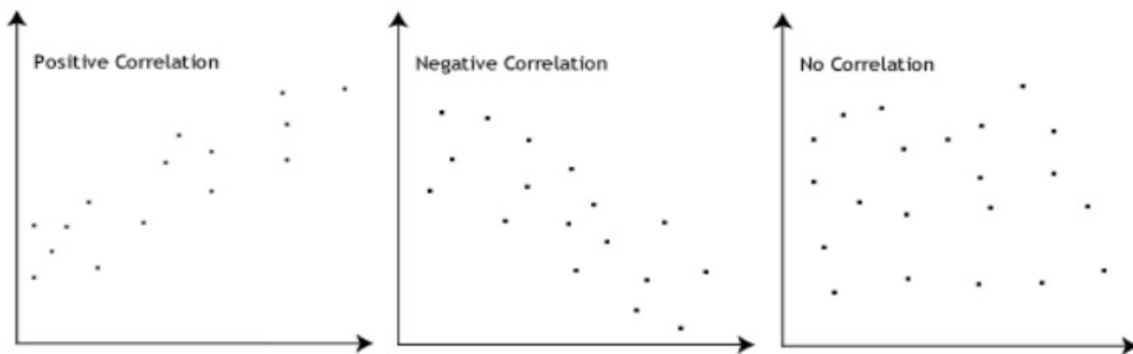
- Outliers Impact: Outliers can significantly affect regression models and correlations, highlighting the need to carefully inspect and handle them.
- Model Fit: The appropriateness of a model cannot be determined by summary statistics alone. It is essential to visualize data to ensure the chosen model fits well. By understanding and applying the lessons from Anscombe's quartet, analysts can avoid common pitfalls in data analysis and ensure more accurate and insightful conclusions.

3. What is Pearson's R?

Answer:

Pearson's R is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, R , can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

This is shown in the diagram below:



The formula of Pearson's R is:

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

Where:

- x_i and y_i are the individual sample points.
- \bar{x} and \bar{y} are the means of the x and y variables, respectively.

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

Answer:

Scaling is the process of adjusting the range of feature values in a dataset so that they fit within a specific range, often to improve the performance and training stability of machine learning algorithms. Why Scaling is Performed

- Improves Model Performance: Many machine learning algorithms perform better when features are on a similar scale. For example, gradient descent converges faster when features are scaled.
- Enhances Model Interpretability: Scaling can help make the features more interpretable and comparable, especially when they are on different scales.
- Prevents Bias Towards Features: Algorithms that compute distances between data points (e.g., k-NN, SVM) or that involve regularization (e.g., linear regression with L2 regularization) can be biased by the scale of features.

Key differences between Normalisation and Standardisation:

Aspect	Normalized Scaling	Standardized Scaling
Formula	$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$	$x' = \frac{x - \mu}{\sigma}$
Range	Typically [0, 1] or [-1, 1]	Mean = 0, Standard Deviation = 1
Sensitivity to Outliers	High	Moderate
Use Case	Non-Gaussian distributions, bounded range	Gaussian distributions, many ML algorithms
Effect	Rescales data to a fixed range	Centers data around mean with unit variance

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

Answer:

If there is perfect correlation, then $VIF = \text{infinity}$. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R^2) =1, which lead to $1/(1-R^2)$ infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

$$VIF(X_i) = \frac{1}{1-R_i^2}$$

Where R_i^2 is the coefficient of determination of the regression of X_i on all other predictors. If $R_i^2 = 1$ (indicating perfect multicollinearity), then:

$$VIF(X_i) = \frac{1}{1-1} = \frac{1}{0} = \infty$$

Causes of Perfect Multicollinearity:

- **Duplicate Variables:** Including the same variable more than once in the regression model.
- **Linear Dependence:** One variable is a perfect linear function of another, such as:

Sum of variables (e.g., $X_3 = X_1 + X_2$).

Constant multiples (e.g., $X_2 = 2X_1$)

- **Dummy Variable Trap:** Including all dummy variables for a categorical feature without dropping one (when using one-hot encoding), causing perfect multicollinearity.

Handling Infinite VIF:

- Remove One of the Collinear Variables: Identify and remove one of the perfectly collinear variables.
- Combine Collinear Variables: Combine collinear variables into a single feature if they convey the same information.

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

Answer:

The quantile-quantile (Q-Q) plot is a graphical method used to determine if two data sets originate from populations with the same distribution.

Use of Q-Q Plot:

A Q-Q plot compares the quantiles of one data set against the quantiles of another. A quantile represents the fraction (or percentage) of points below a specific value. For example, the 0.3 (or 30%) quantile is the value below which 30% of the data falls. The plot also includes a 45-degree reference line. If both data sets come from the same distribution, the points will align closely with this line. Deviations from the line suggest differences in the distributions.

Importance of Q-Q Plot:

When comparing two data samples, it's important to know if they share a common distribution. If they do, estimates of location and scale can be pooled from both samples for more accurate results. If the samples differ, the Q-Q plot helps to understand the nature of these differences better than some analytical methods, such as the chi-square or Kolmogorov-Smirnov tests.