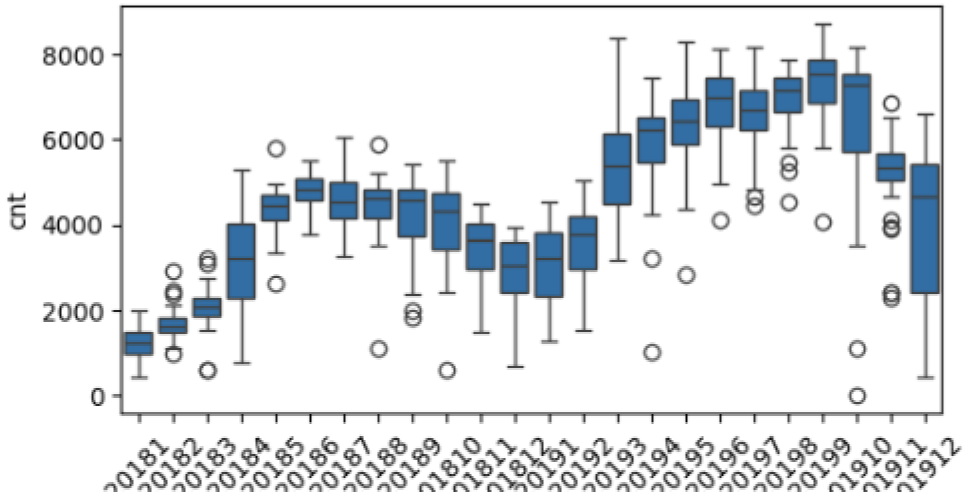
Linear Regression Subjective Questions

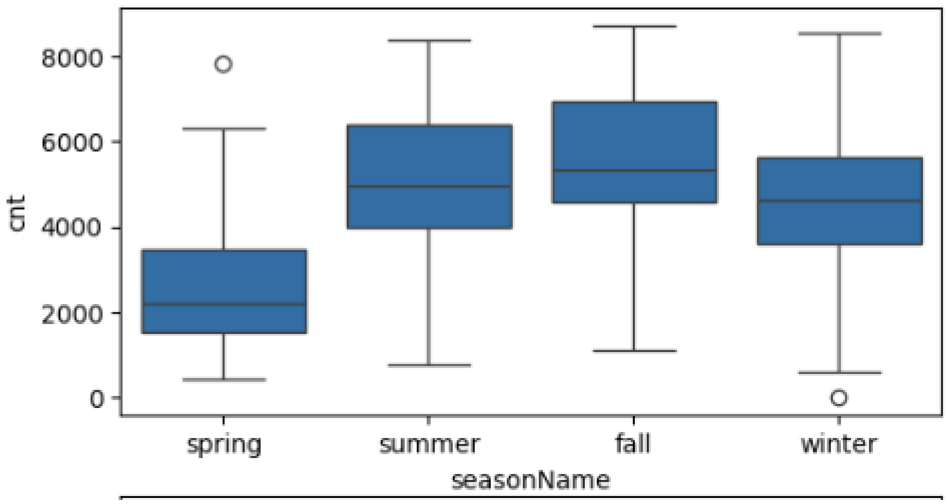
Assignment-based Subjective Questions

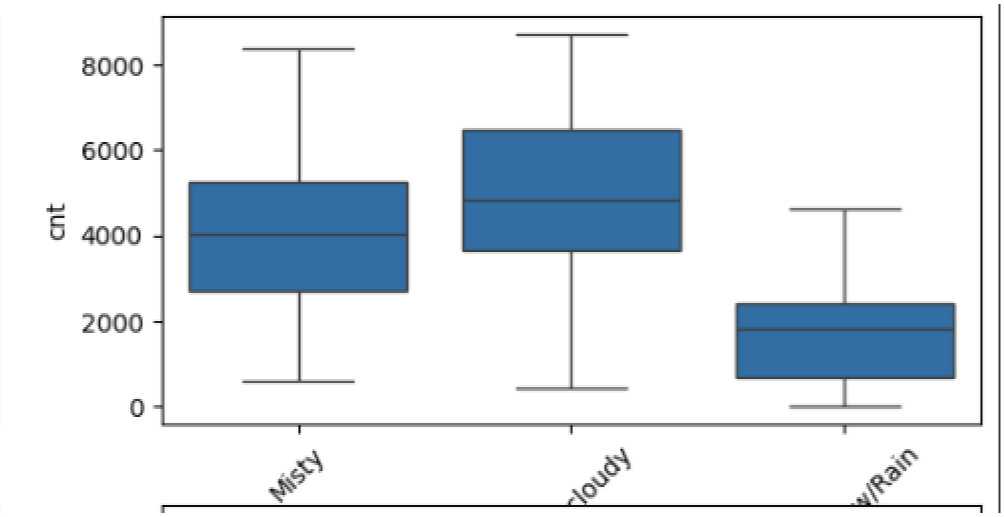
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

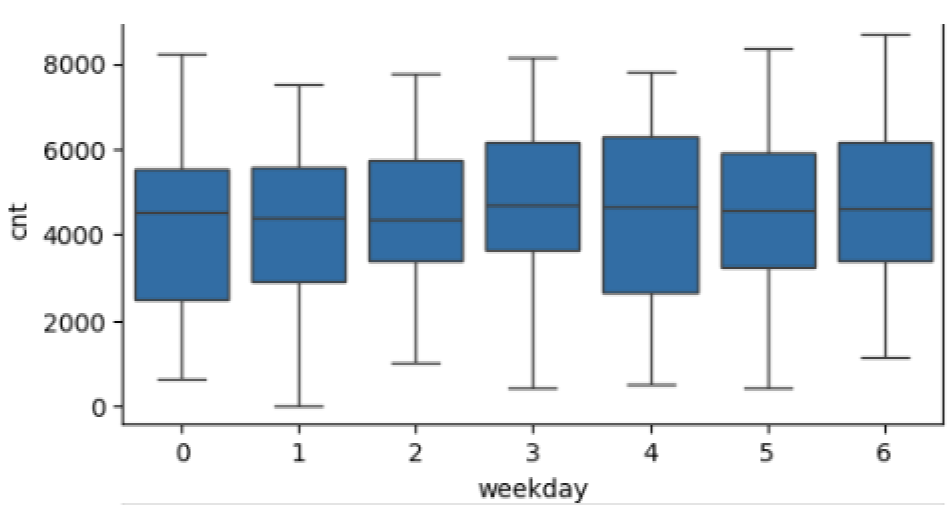
- For the given dataset we have categorical variables year, month, weekday, `season`, `weathersit.`

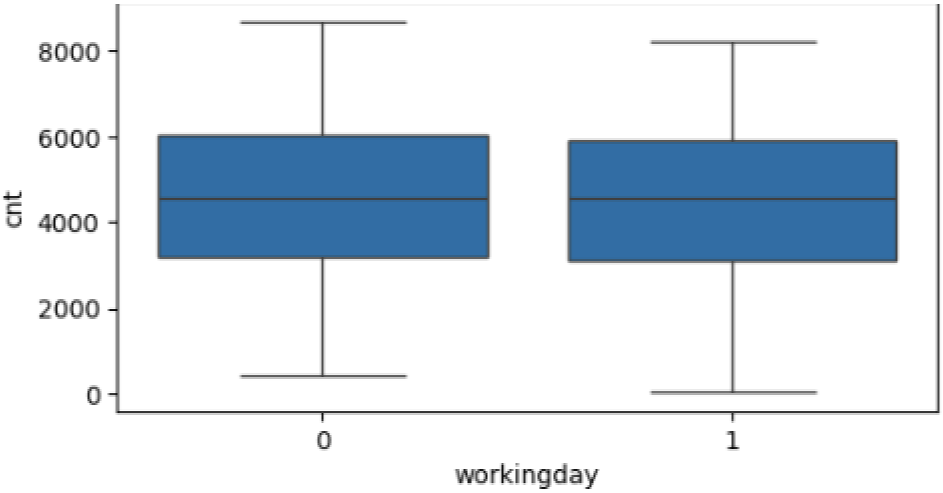
```
From the below box plots we could clearly see except for workingday all
other have some effect on the dependent variable
```

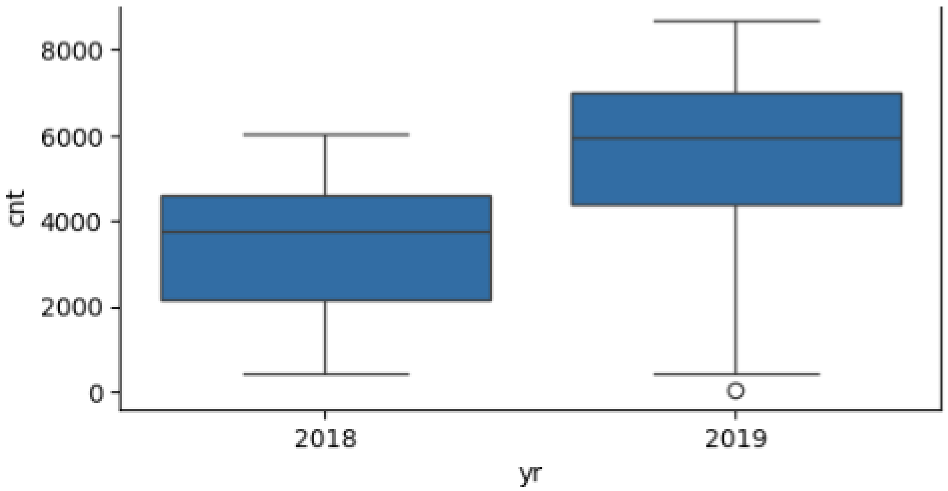


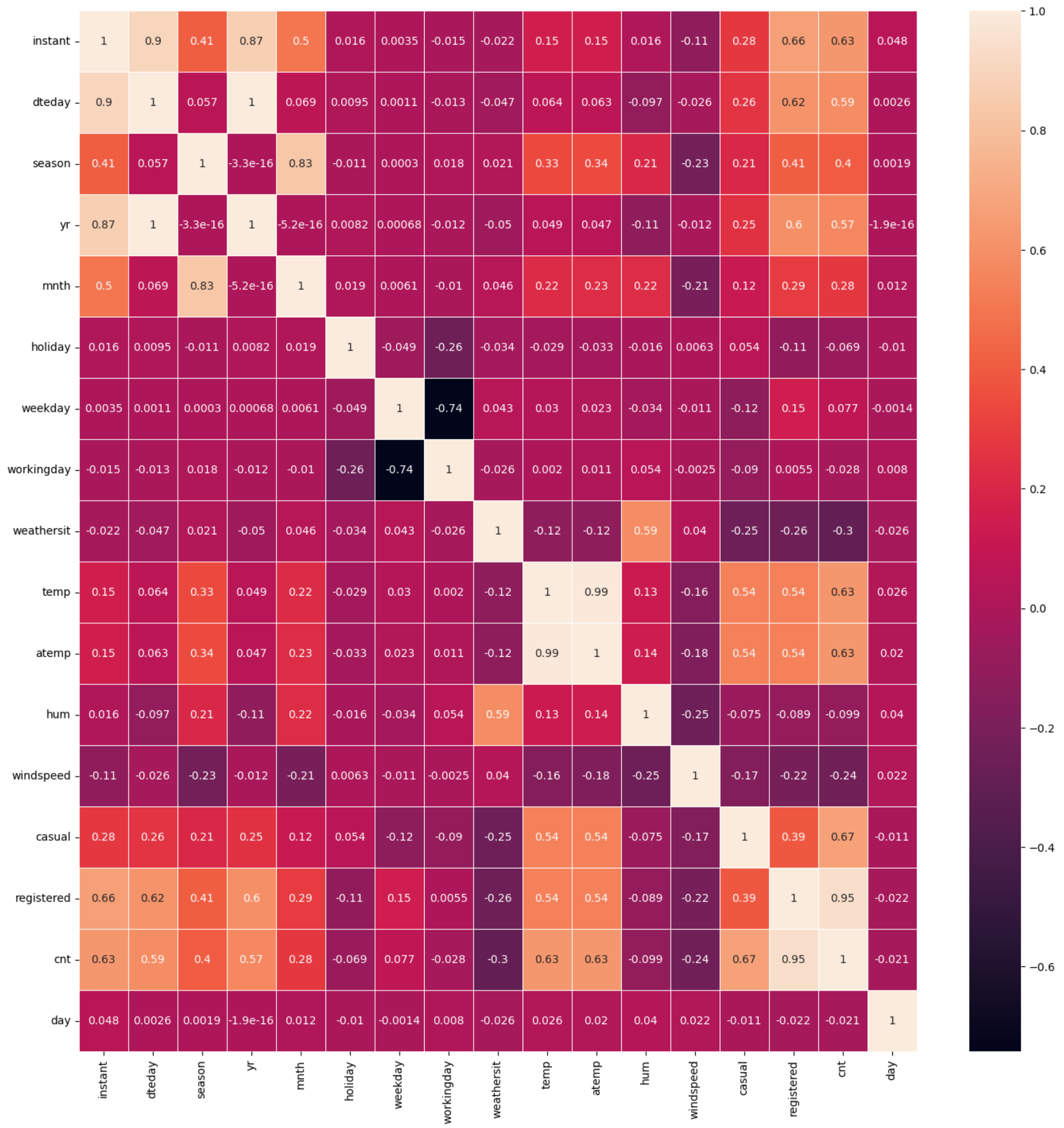



8000
6000
4000
2000
0
cnt
Misty
cloudy
w/Rain


8000
6000
4000
2000
0
cnt
0
1
2
3
4
5
6
weekday


8000
6000
4000
2000
0
cnt
0
1
workingday


8000
6000
4000
2000
0
cnt
2018
2019
yr



	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt	day
instant	1	0.9	0.41	0.87	0.5	0.016	0.0035	-0.015	-0.022	0.15	0.15	0.016	-0.11	0.28	0.66	0.63	0.048
dteday	0.9	1	0.057	1	0.069	0.0095	0.0011	-0.013	-0.047	0.064	0.063	-0.097	-0.026	0.26	0.62	0.59	0.0026
season	0.41	0.057	1	-3.3e-16	0.83	-0.011	0.0003	0.018	0.021	0.33	0.34	0.21	-0.23	0.21	0.41	0.4	0.0019
yr	0.87	1	-3.3e-16	1	-5.2e-16	0.0082	0.00068	-0.012	-0.05	0.049	0.047	-0.11	-0.012	0.25	0.6	0.57	-1.9e-16
mnth	0.5	0.069	0.83	-5.2e-16	1	0.019	0.0061	-0.01	0.046	0.22	0.23	0.22	-0.21	0.12	0.29	0.28	0.012
holiday	0.016	0.0095	-0.011	0.0082	0.019	1	-0.049	-0.26	-0.034	-0.029	-0.033	-0.016	0.0063	0.054	-0.11	-0.069	-0.01
weekday	0.0035	0.0011	0.0003	0.00068	0.0061	-0.049	1	-0.74	0.043	0.03	0.023	-0.034	-0.011	-0.12	0.15	0.077	-0.0014
workingday	-0.015	-0.013	0.018	-0.012	-0.01	-0.26	-0.74	1	-0.026	0.002	0.011	0.054	-0.0025	-0.09	0.0055	-0.028	0.008
weathersit	-0.022	-0.047	0.021	-0.05	0.046	-0.034	0.043	-0.026	1	-0.12	-0.12	0.59	0.04	-0.25	-0.26	-0.3	-0.026
temp	0.15	0.064	0.33	0.049	0.22	-0.029	0.03	0.002	-0.12	1	0.99	0.13	-0.16	0.54	0.54	0.63	0.026
atemp	0.15	0.063	0.34	0.047	0.23	-0.033	0.023	0.011	-0.12	0.99	1	0.14	-0.18	0.54	0.54	0.63	0.02
hum	0.016	-0.097	0.21	-0.11	0.22	-0.016	-0.034	0.054	0.59	0.13	0.14	1	-0.25	-0.075	-0.089	-0.099	0.04
windspeed	-0.11	-0.026	-0.23	-0.012	-0.21	0.0063	-0.011	-0.0025	0.04	-0.16	-0.18	-0.25	1	-0.17	-0.22	-0.24	0.022
casual	0.28	0.26	0.21	0.25	0.12	0.054	-0.12	-0.09	-0.25	0.54	0.54	-0.075	-0.17	1	0.39	0.67	-0.011
registered	0.66	0.62	0.41	0.6	0.29	-0.11	0.15	0.0055	-0.26	0.54	0.54	-0.089	-0.22	0.39	1	0.95	-0.022
cnt	0.63	0.59	0.4	0.57	0.28	-0.069	0.077	-0.028	-0.3	0.63	0.63	-0.099	-0.24	0.67	0.95	1	-0.021
day	0.048	0.0026	0.0019	-1.9e-16	0.012	-0.01	-0.0014	0.008	-0.026	0.026	0.02	0.04	0.022	-0.011	-0.022	-0.021	1

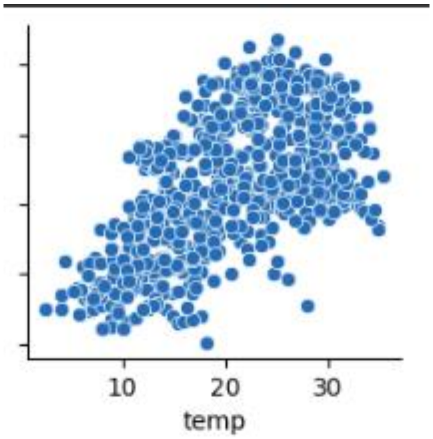
From the heatmap we could see temperature has the highest impact(0.63) on the derived variable(cnt), followed by yr(corona & non-corona period) impact(0.57), followed by seasons, weather situation.

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

- To have the values range between 0 and 1 we want to create dummy variables, if we have n different values in the category we could identify each value by creating n-1 dummy variables. Here to drop that one column we use argument drop_first=True in pandas get_dummies function, which will remove the first column.

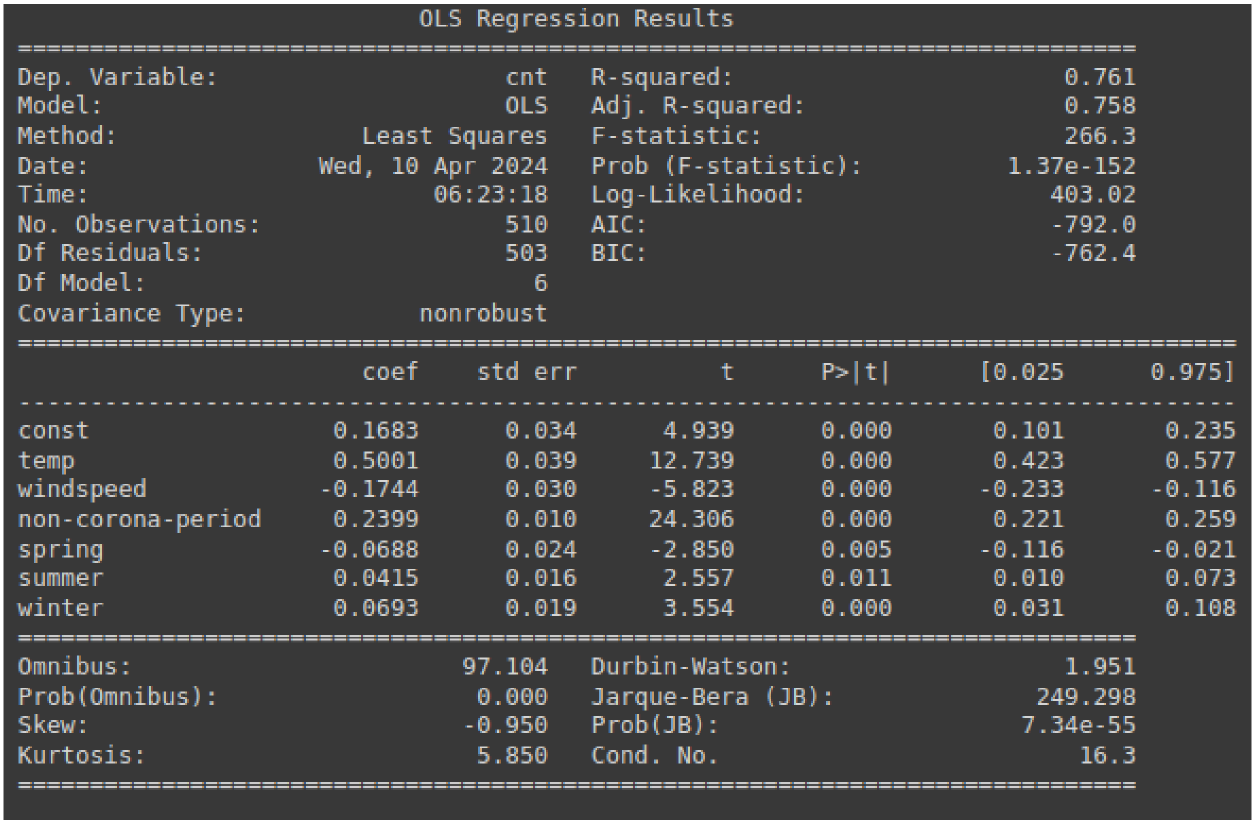
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

- Temperature has the highest correlation



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

- I did a data split of available data into a 70-30 ratio for training and test purpose. On the training data I evaluated the linera regression model summary and verified that the p-values are zero’s or almost zero’s. Followed by checking the R-squared is a decent value explaining the prediction, Prob (F-statistic) is almost zero.
- I ran the model on the test data to verify the R2 value of train data and test data almost match.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.761
Model:                            OLS   Adj. R-squared:                  0.758
Method:                 Least Squares   F-statistic:                     266.3
Date:                Wed, 10 Apr 2024   Prob (F-statistic):          1.37e-152
Time:                        06:23:18   Log-Likelihood:                 403.02
No. Observations:                 510   AIC:                            -792.0
Df Residuals:                     503   BIC:                            -762.4
Df Model:                           6
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.1683      0.034      4.939      0.000       0.101       0.235
temp                  0.5001      0.039     12.739      0.000       0.423       0.577
windspeed            -0.1744      0.030     -5.823      0.000      -0.233      -0.116
non-corona-period     0.2399      0.010     24.306      0.000       0.221       0.259
spring               -0.0688      0.024     -2.850      0.005      -0.116      -0.021
summer                0.0415      0.016      2.557      0.011       0.010       0.073
winter                0.0693      0.019      3.554      0.000       0.031       0.108
==============================================================================
Omnibus:                       97.104   Durbin-Watson:                   1.951
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              249.298
Skew:                          -0.950   Prob(JB):                     7.34e-55
Kurtosis:                       5.850   Cond. No.                         16.3
==============================================================================
```

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- Temperature
- non-corona-period(year)
- Season

General Subjective Questions

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

- Linear regression is a fundamental supervised machine learning algorithm used for predicting a continuous target variable based on one or more input features. It models the relationship between the independent variables (features) and the dependent variable (target) by fitting a linear equation to the observed data.

  The objective is to find the optimal values for the coefficients that minimize the sum of squared differences between the observed and predicted values. This is typically done using Ordinary Least Squares (OLS) or other optimization techniques. Once the coefficients are estimated, predictions can be made for new input values. The model's performance can be evaluated using metrics like Mean Squared Error (MSE) or R-squared. Linear regression assumes several assumptions about the data, including linearity, independence of errors, homoscedasticity, and normality of residuals. It can be extended to handle more complex scenarios and regularization techniques can be applied to prevent overfitting.

Linear regression assumes that there is a linear relationship between the independent variables X and the dependent variable Y. Mathematically, it can be represented as:

$Y=\beta_0 +\beta_1 X_1 +\beta_2 X_2 +\ldots+\beta_n X_n +\varepsilon$

Where:

Y is the dependent variable (target).

$X_1 ,X_2 ,\ldots,X_n$ are the independent variables (features).

β_0 is the y-intercept (constant term).

$\beta_1 ,\beta_2 ,\ldots,\beta_n$ are the coefficients (slope).

ε is the error term (residuals), representing the difference between the observed and predicted values.

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

Anscombe's quartet comprises four datasets with nearly identical statistical properties but distinctly different visual representations. Proposed by statistician Francis Anscombe in 1973, this quartet serves to underscore the critical importance of visualizing data rather than relying solely on summary statistics.

Description of the Quartet:

Each dataset in Anscombe's quartet consists of 11 data points. Despite their similar summary statistics, such as means, variances, and correlations, they display diverse patterns when graphed.

Properties of the Quartet:

Dataset I: Exhibits a linear relationship with some random noise.

Dataset II: Features a linear relationship with an outlier that notably influences linear regression.

Dataset III: Demonstrates a non-linear relationship well-suited for a quadratic regression model.

Dataset IV: Shows an apparent relationship except for one outlier, which significantly impacts the correlation coefficient.

Implications:

This quartet emphasizes that summary statistics alone may not fully encapsulate the essence of the data. Key takeaways include:

Similar summary statistics across datasets can belie vastly different underlying patterns.

Visualization plays a pivotal role in unraveling data characteristics.

Outliers can markedly skew statistical analyses and modeling outcomes.

Importance:

Anscombe's quartet underscores the indispensable role of data visualization in exploratory data analysis and model interpretation. It serves as a cautionary example, warning against the pitfalls of relying solely on summary statistics. Instead, analysts should embrace comprehensive data exploration and visualization to make well-informed decisions.

Practical Applications:

In real-world scenarios, Anscombe's quartet serves as a potent reminder for statisticians and data analysts. It advocates for thorough data exploration and visualization before drawing conclusions or building models based solely on summary statistics. Often utilized in statistics education, it elucidates the significance of graphical analysis and the limitations of summary statistics.

In essence, Anscombe's quartet compellingly challenges the overreliance on summary statistics, advocating for a balanced approach that integrates visual exploration for a deeper understanding of data dynamics.

3. What is Pearson’s R? (3 marks)

Pearson's correlation coefficient, also referred to as Pearson's r, is a statistical metric that measures the degree and direction of the linear relationship between two continuous variables. It was introduced by Karl Pearson in the late 19th century and finds widespread use across disciplines such as statistics, psychology, economics, and social sciences.

Key Features:

Range: Pearson's r spans from -1 to +1.

Direction:

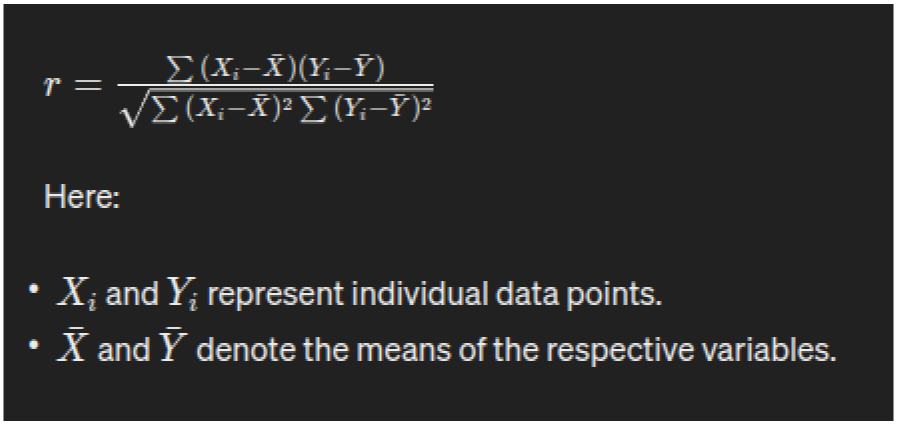
A positive value (r>0) signifies a positive linear relationship, where an increase in one variable tends to accompany an increase in the other.

Conversely, a negative value (r<0) indicates a negative linear relationship, suggesting that as one variable rises, the other tends to decline.

Strength: The magnitude of r reflects the strength of the linear relationship. Values closer to +1 or -1 denote stronger correlations, while 0 signifies no linear relationship.

Formula:

Pearson's correlation coefficient (r) is computed using the following formula:



$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Here:

- X_i and Y_i represent individual data points.
- $\bar{X}$ and $\bar{Y}$ denote the means of the respective variables.

Interpretation:

r=1 signifies a perfect positive linear relationship.

r=−1 indicates a perfect negative linear relationship.

r=0 denotes no discernible linear relationship.

Values of r between 0 and ±1 portray the strength and direction of the linear relationship.

Assumptions and Constraints:

Pearson's r assumes linearity between the variables.

It's sensitive to outliers and may not adequately capture non-linear relationships.

This metric solely gauges the strength and direction of linear associations, without implying causation.

Application:

Pearson's correlation coefficient is frequently employed in data analysis to:

Assess the connection between two continuous variables.

Detect trends or patterns in datasets.

Evaluate the reliability and validity of measurement tools.

Inform decision-making across a spectrum of fields, including research, business, and healthcare.

In essence, Pearson's correlation coefficient serves as a valuable statistical tool for gauging the degree and direction of linear associations between continuous variables, thus facilitating deeper insights into data relationships.

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

Scaling is a preprocessing technique used in data analysis and machine learning to transform the features of a dataset to a similar scale. This process involves adjusting the values of the features to fall within a specified range, typically between 0 and 1 or with a mean of 0 and a standard deviation of 1. Scaling is performed to ensure that all features contribute equally to the analysis, prevent features with larger scales from dominating, and to improve the performance of certain algorithms.

Why Scaling is Performed:

Equal Contribution: Scaling ensures that all features contribute equally to the analysis. Features with larger scales can otherwise dominate the analysis, leading to biased results.

Algorithm Sensitivity: Some machine learning algorithms are sensitive to the scale of the features. Scaling helps algorithms converge faster and prevents them from getting stuck in local optima.

Distance-based Algorithms: Algorithms that use distances between data points, such as k-nearest neighbors (KNN) and support vector machines (SVM), are sensitive to feature scales. Scaling ensures that distances are calculated accurately.

Regularization: Regularization techniques, like ridge regression and lasso regression, penalize large coefficients. Scaling helps prevent features with larger scales from receiving disproportionate penalties.

Difference between Normalized Scaling and Standardized Scaling:

Normalized Scaling (Min-Max Scaling):

> Range: Adjusts the values to fall within a specified range, usually between 0 and 1.
>
> Formula: Xscaled = (Xmax −Xmin) / (X−Xmin)
>
> Pros: Simple to implement, maintains the shape of the original distribution.
>
> Cons: Sensitive to outliers, does not handle outliers well.

Standardized Scaling (Z-score Scaling):

> Mean and Standard Deviation: Adjusts the values to have a mean of 0 and a standard deviation of 1.
>
> Formula: Xscaled =σX−μ , where μ is the mean and σ is the standard deviation.
>
> Pros: Robust to outliers, maintains the relative relationships between data points.
>
> Cons: Does not bound the data to a specific range.

Summary:

Normalized scaling adjusts the values to fall within a specific range (usually 0 to 1), while standardized scaling adjusts the values to have a mean of 0 and a standard deviation of 1.

Normalized scaling is simple and maintains the shape of the original distribution but is sensitive to outliers. Standardized scaling is robust to outliers but does not bound the data to a specific range.

The choice between the two scaling methods depends on the specific requirements of the analysis and the characteristics of the dataset.

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

   Yes, occasionally the Variance Inflation Factor (VIF) can yield an infinite value. This occurs when the correlation between one or more predictor variables is perfect or extremely high, leading to numerical instability in the calculation of VIF.

   Explanation:

   Perfect Multicollinearity: When two or more predictor variables are perfectly correlated, it means that one predictor variable can be expressed as a linear combination of the others. In such cases, the matrix used to calculate VIF becomes singular, resulting in a division by zero error and an infinite VIF value.

   Extremely High Correlation: Even if multicollinearity is not perfect but extremely high, it can still cause numerical instability in the VIF calculation. This can happen when the correlation coefficient between predictor variables is very close to 1.

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

   A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the dataset's empirical distribution (observed data) against the quantiles of a specified theoretical distribution (expected data). The Q-Q plot helps to visually inspect how closely the observed data match the theoretical distribution, enabling analysts to assess assumptions about the data's distributional shape.

   Q-Q plots are valuable tools in linear regression analysis for evaluating the normality assumption of residuals. They enable analysts to detect departures from normality, validate regression models, compare models, and diagnose potential issues, ultimately improving the accuracy and reliability of regression analysis results.