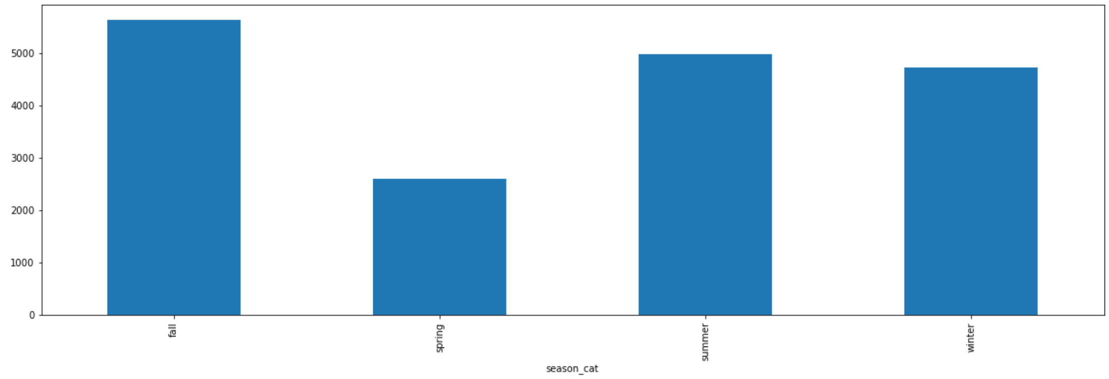
Assignment-based Subjective Questions

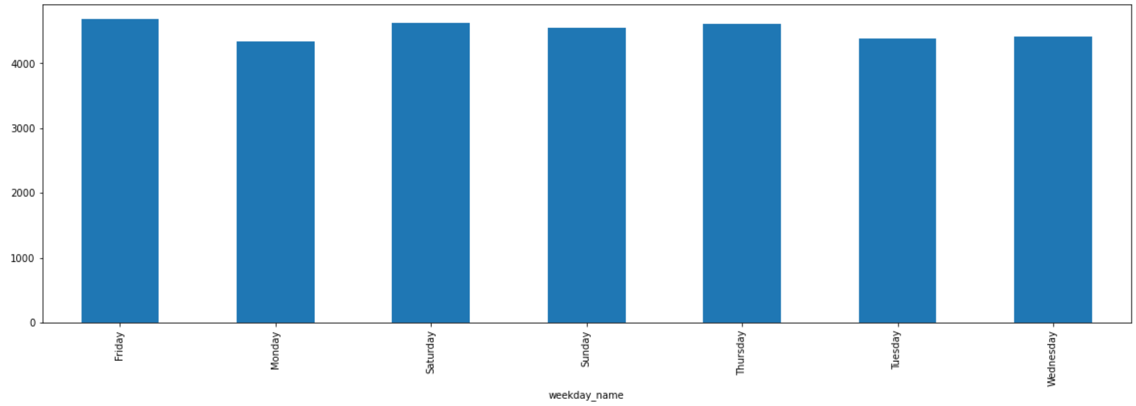
From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Below are the inferences for the categorical variables -

1. Bike demand mean is less in Spring and high in Fall.

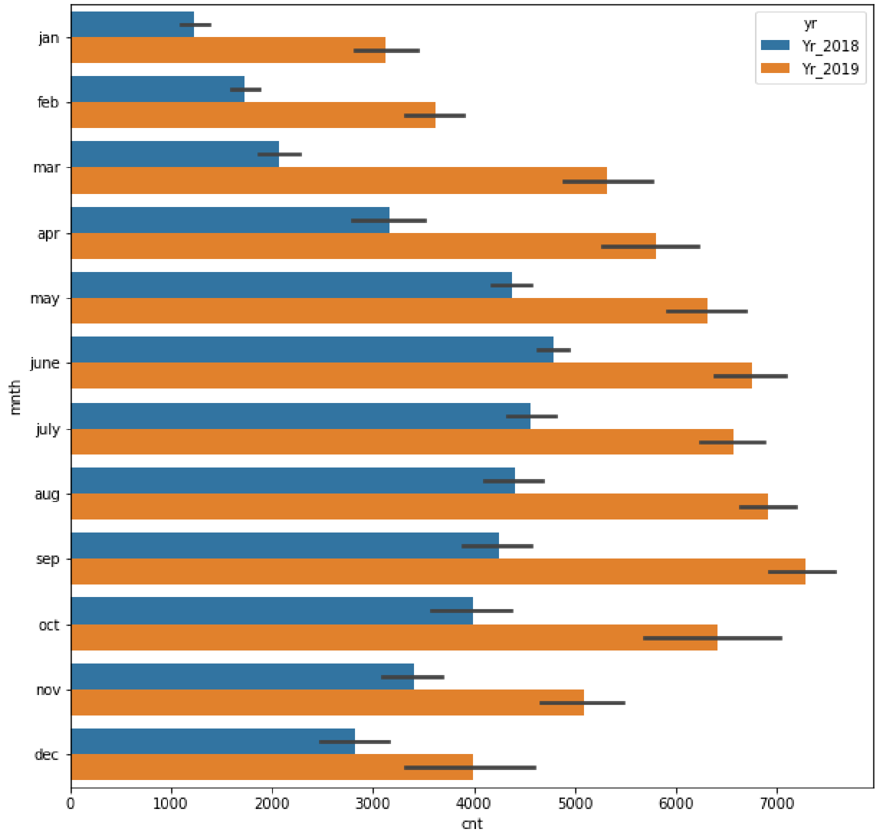


2. Day in particular doesn't have much mean difference on Bike Demand

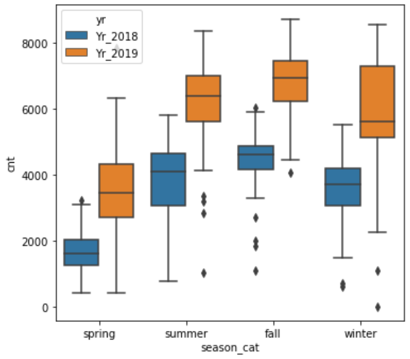


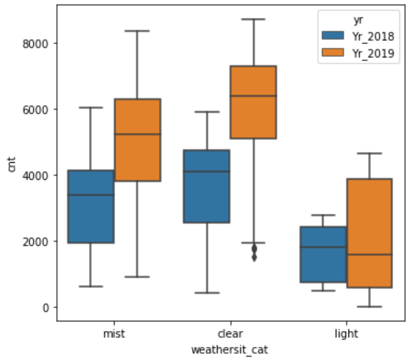
3.

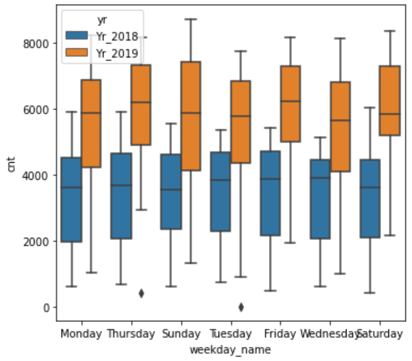
Demand is seen reducing in July month – in 2018 and 2019

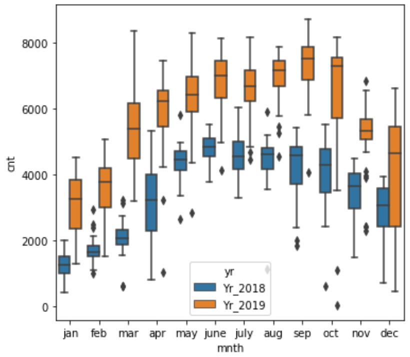


Additional Analysis -









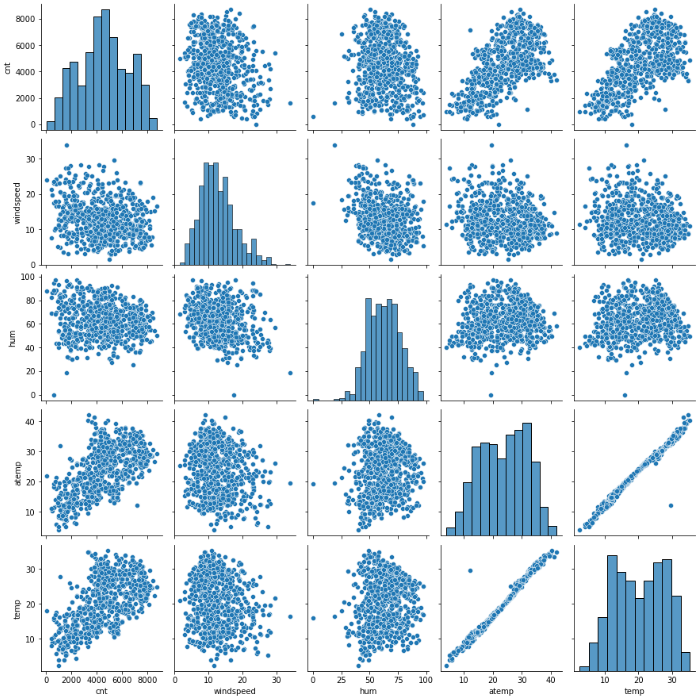
Why is it important to use drop_first=True during dummy variable creation?

Which creating dummy variables, by default get_dummies create bool variable for all the levels in the given variable. If a variable has n levels, for analysis it is sufficient to have n-1 variables. **It will reduce the unnecessary variable creation**.

Eg: Variable name: Check, it has values of “Yes”, “No”, “Maybe”. For this example, it is enough to have two bool variables to understand the three levels.

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

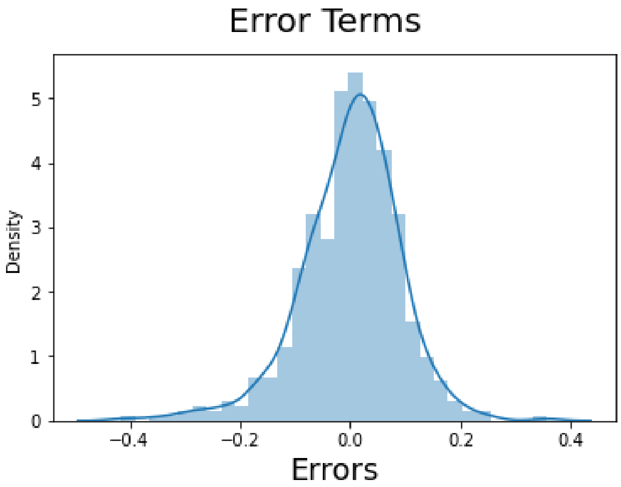
Answer: Both, Temperature and Feeling Temperature having good correlation with target variable – cnt.



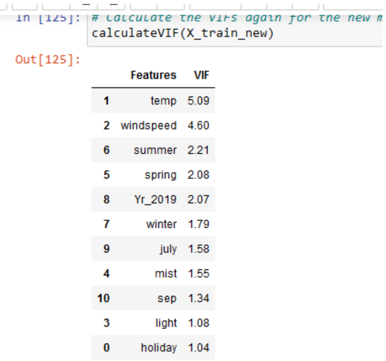
How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

1. Error Terms are normally distributed:



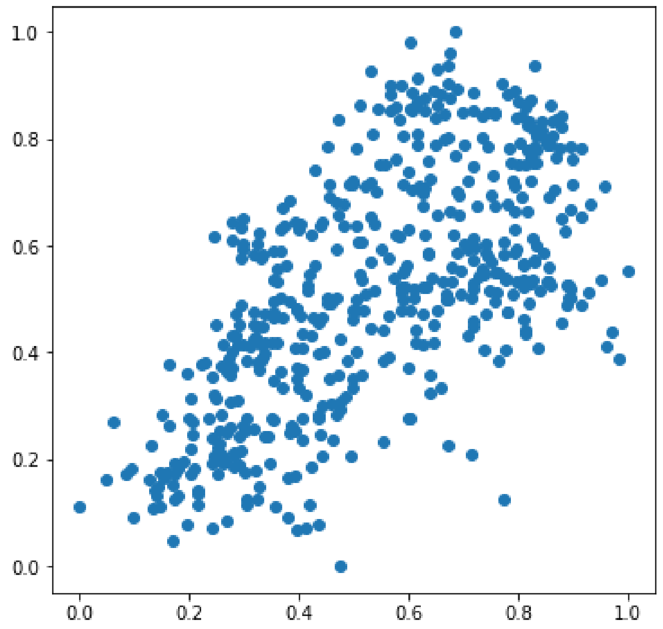
2. Multicollinearity checked:



```
In [125]: # calculate the VIFs again for the new m
          calculateVIF(X_train_new)
Out[125]:
```

	Features	VIF
1	temp	5.09
2	windspeed	4.60
6	summer	2.21
5	spring	2.08
8	Yr_2019	2.07
7	winter	1.79
9	july	1.58
4	mist	1.55
10	sep	1.34
3	light	1.08
0	holiday	1.04

3. Linera relationship in predictor and target variable seen. Temp vs Cnt



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

Answer:

Temp, Winter and September month seen positive correlation

Holiday, Windspeed and Light Rain seen negative correlation

Bike_Demand = 0.1996 + temp * 0.4915 + Yr_2019 * 0.2335 + winter * 0.0831 +
sep * 0.0767 + summer * 0.0453 - july * 0.0524 - spring * 0.0669 -
mist * 0.0816 - holiday * 0.098 - windspeed * 0.148 - light * 0.2852

General Subjective Questions

Explain the linear regression algorithm in detail.

Answer:

Linear Regression (LR) is a process of estimating relationship between variables. It focuses on the relationship between dependent (Target) and independent variables (Predictors).

It explains the change in dependent variable with change in the values of predictor.

There are two types of LR,

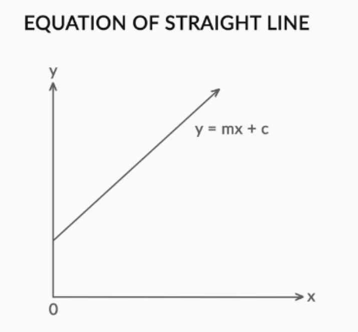
1. Simple linear regression, where there is a change in only one predictor at a time
2. Multiple linear regression, where there is change in multiple variables

LR describes the Target (y) and Predictor (x) using Stright line, as below -

$$y = mx + c$$

Where -

- "y" is the Target
- "x" is the predictor
- "m" is the slope
- "c" is the intercept



The strength of a linear regression line is determined by the value of the **R-squared**. The value of R-squared lies between 0 and 1, where 1 implies that the variance in the data is being explained by the model, and 0 implies that none of the variance values is being explained by the model.

Explain the Anscombe's quartet in detail.

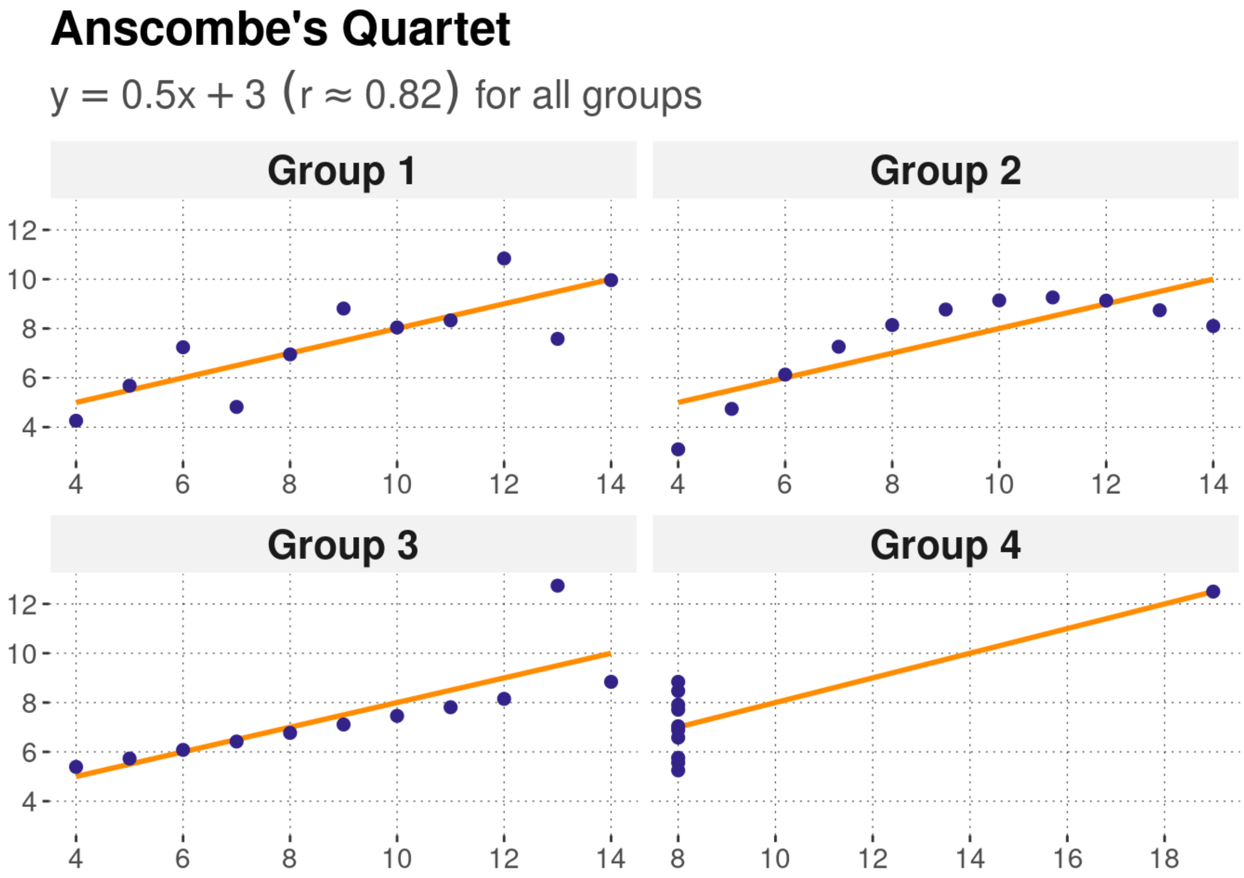
Answer:

Anscobe’s quartet is set of 4 datasets that shares nearly identical simple descriptive statistics. But exhibits vastly different distributions when graphed, as we can see in the screenshot below.

- same mean

- same standard deviation

- same regression line



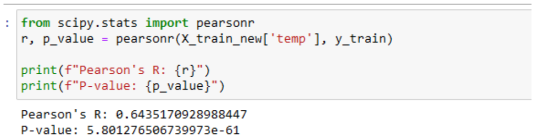
It highlights the impact of outliers and observations on statistical properties.

What is Pearson’s R?

Answer:

Pearson's R, also known as the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables

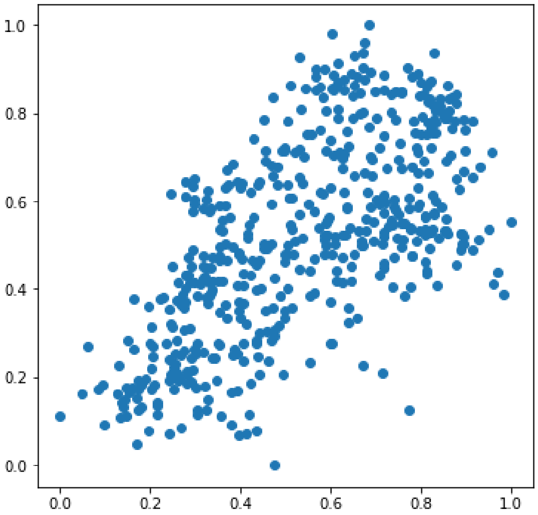
Pearson’s R value(p-value), confirming the strength and direction of the linear relationship between the two variables. In the Assignment we can see the positive correlation with Temp predictor and Target Cnt.



```
from scipy.stats import pearsonr
r, p_value = pearsonr(X_train_new['temp'], y_train)

print(f"Pearson's R: {r}")
print(f"P-value: {p_value}")
```

```
Pearson's R: 0.6435170928988447
P-value: 5.801276506739973e-61
```



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

- Scaling is a data pre-processing technique used to adjust the range of feature values. **It only affects** the coefficients and other attributes will not be impacted such as T-Stat,F-Stat,P-Value,R2 etc.
- Scaling performed for ease of interpretation and faster conversance.
- Standardization and Normalization (Min-Max) are two types Scaling:
    - Nomalization (Min-Max) :
        - Adjusts the feature values between 0 and 1. The formula: (X−Xmin) / (Xmax −Xmin), X is original feature value
        - Outlier as well try to fit with in 0 and 1
    - Standardization:

- Adjusts the values of a feature to have a mean of 0 and a standard deviation of 1. The formula: (X-Mean) / (std deviation)
- Outlier considered in mean and std dev

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

Answer:

VIF – Variance Inflation Factor : is used to detect the associations in predictors, explain the correlation.

VIF is calculated using R^2 value. As (1-R^2) $^{-1}$

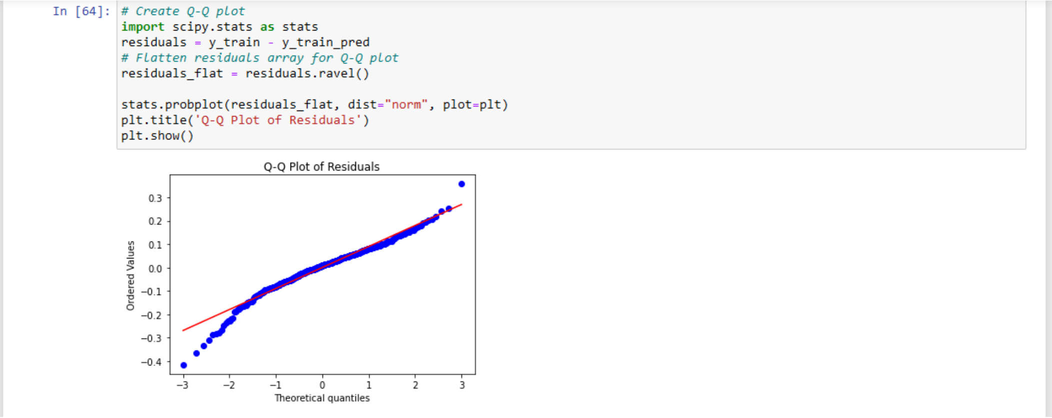
When R^2 value becomes 1 (100%) then the Value of VIF becomes infinite. Which means both the predictors are perfectly corelated – perfect linear combination.

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

Answer:

A Q-Q plot, or Quantile-Quantile plot, used to check the assumption of normality of the residuals.

Normality of Residuals: One of the key assumptions in linear regression is that the residuals (errors) are normally distributed. The Q-Q plot helps visually assess this assumption. Below is what we have for Bike demand dataset.



```
In [64]: # Create Q-Q plot
         import scipy.stats as stats
         residuals = y_train - y_train_pred
         # Flatten residuals array for Q-Q plot
         residuals_flat = residuals.ravel()

         stats.probplot(residuals_flat, dist="norm", plot=plt)
         plt.title('Q-Q Plot of Residuals')
         plt.show()
```