1. From your analysis of the categorical variables from the dataset, what could you infer about

their effect on the dependent variable?

Insights

There were 6 categorical variables in the dataset.

We used Box plot (refer the fig above) to study their effect on the dependent variable (‘cnt’) .

The inference that We could derive were:

* **season**: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.
* **mnth**: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
* **weathersit**: Almost 67% of the bike booking were happening during ‘weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
* **holiday**: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
* **weekday**: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.
* **workingday**: Almost 69% of the bike booking were happening in ‘workingday’ with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

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

* drop\_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
* Let’s say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi\_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished. [Example](https://i.stack.imgur.com/nMWkQ.png)

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

3)Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable?

* The above Pair-Plot tells us that there is a LINEAR RELATION between 'temp','atemp' and 'cnt'

4)How did you validate the assumptions of Linear Regression after building the model on the

training set?

By looking at the scatter plot

Scatter plot between y\_test and y\_pred mimics linear relation

Also when we perform residual analysis on the train set

*# Mean of Residuals*

(y\_train-y\_train\_pred).mean()

We can clearly observe that

* Residual errors follow a normal distribution with mean=0
* Variance of Errors doesnt follow any trends
* Residual errors are independent of each other since the Predicted values vs Residuals plot doesn't show any trend.

5) Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

As per our final Model, the top 3 predictor variables that influences the bike booking are:

* **Temperature (temp)** - A coefficient value of ‘0.5636’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.
* **Weather Situation 3 (weathersit\_3)** - A coefficient value of ‘-0.3070’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units.
* **Year (yr)** - A coefficient value of ‘0.2308’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.