Abi Glisson

ENTITY Data Science Virtual Academy – Final Project

8/12/22

Analysis Planning Worksheet

# Evaluation Question

Prediction of bike rental count hourly based on environmental and seasonal settings *(This should end up as a multiple linear regression using supervised machine learning)*

# Independent Variable(s)

These variable(s) are causing something or creating an effect. List what each is and whether it is categorical or continuous. It is ok to only have one.

**Variable**

`season` - seasons of the year – 4 levels: 1: winter, 2: spring, 3: summer, 4: fall

Categorical: # of levels \_\_4\_\_\_  Continuous

**Variable**

`yr` - year – 0: 2011, 1: 2012

Categorical: # of levels \_\_2\_\_\_  Continuous

**Variable**

`mnth` - month – 1: January through 12: December

(not 100% sure whether this is categorical or continuous. December isn’t actually “bigger” than January, and we repeat each year, so I think it’s categorical?)

Categorical: # of levels \_\_12\_\_\_  Continuous

Variable

`hr` - hour – 0: 12:00AM to 23: 11:00PM

Categorical: # of levels \_\_24\_\_\_  Continuous

Variable

`holiday` - 0: not a holiday, 1: a holiday

Categorical: # of levels \_\_2\_\_\_  Continuous

Variable

`weekday` - day of the week – 0: Sunday through 6: Saturday

Categorical: # of levels \_\_7\_\_\_  Continuous

Variable

`workingday` - whether it was a work day – 0: weekend or holiday, 1: not weekend or holiday

Categorical: # of levels \_\_2\_\_\_  Continuous

Variable

`weathersit` - weather situation (?) - weathersit

1: Clear, Few clouds, Partly cloudy, Partly cloudy

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

Categorical: # of levels \_\_4\_\_\_  Continuous

Variable

`temp` - Normalized temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 (only in hourly scale)

Categorical: # of levels \_\_\_\_\_  Continuous

Variable

`atemp` - Normalized feeling temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-16, t\_max=+50 (only in hourly scale)

Categorical: # of levels \_\_\_\_\_  Continuous

Variable

`hum` - Normalized humidity. The values are divided to 100 (max)

Categorical: # of levels \_\_\_\_\_  Continuous

Variable

`windspeed` - Normalized wind speed. The values are divided to 67 (max)

Categorical: # of levels \_\_\_\_\_  Continuous

# Dependent Variable(s)

These variable(s) are influenced by your independent variable and *depend* on them. List what each is and whether it is categorical or continuous. Unless they are related, you should have only one.

**Variable**

`casual` - count of casual users

Categorical: # of levels \_\_\_\_\_  Continuous

**Variable**

`registered`: count of registered users

Categorical: # of levels \_\_\_\_\_  Continuous

**Variable**

`cnt` - count of total rental bikes including both casual and registered

Categorical: # of levels \_\_\_\_\_  Continuous

*The 3 DVs listed above ARE related in that `cnt` = `casual` + `registered`. We will see how this relationship plays out once we get into the analysis.*

Now that you know the type and number of independent and dependent variables, you are ready to use the analysis flow charts to choose your analysis!

# Analysis

Dr. Mo has already told us that we will be performing a multiple linear regression using supervised machine learning in order to predict hourly bike rental counts. I believe that is the case. However, as the multiple linear regression only uses continuous IVs and continuous DVs, I believe there is also an opportunity to perform an ANOVA using the categorical IVs to predict the continuous DV. We may also perform multiple versions of these analyses to determine whether the split of registered vs casual users shows any differences compared to the total number of users across types of days (I believe this will require a type of Chi-Square analysis.