Project 1: NYC Subway Ridership

Section 1. Statistical Test

1.1A: Mann-Whitney U-test

1.1B: Two-tailed

1.1C: There is no difference in NYC Subway ridership if it is raining.

1.1D: .05 or 50%

1.2: The Mann-Whitney U-test is applicable in this instance because it is a non-parametric test that allows both of our conditions (with rain vs. without rain) to be compared without the assumption that values are normally distributed for ENTRIESn_hourly (Turnstile Entries).

1.3: With Rain Mean: 1105.4463767458733

Without Rain Mean: 1090.278780151855

U-Value: 1924409167.0

P-Value: 0.024999912793489721

1.4: Our null hypothesis can be rejected since we have determined our p-value to be 2.49%, far less than the significance level of 50%, paired with information from each mean value indicator. In conclusion, more people ride the NYC subway when it is raining.

Section 2. Linear Regression

2.1: Gradient Descent (As implemented in exercise 3.5)

2.2: Input Variables - Rain / Precipitation / Mean Temperature / Hour

Dummy Variables - UNIT

2.3: I selected these weather related features because I believed they would produce the most accurate representation of ridership differences in a linear model.

2.4: The coefficients (weights) for Rain / Precipitation / Mean Temperature / Hour. The extras show the dummy theta values.

[2.92398062e+00 1.46526720e+01 4.67708502e+02 -6.22179395e+01

1.50048977e+02 -1.51041835e+01 -2.78178403e+01 -1.78111613e+01

-3.82710650e+00 -1.51064914e+01 -3.03405205e+01 -2.81397316e+01

-2.72977045e+01 1.66072856e+02 2.13089119e+02 2.67066331e+02 …

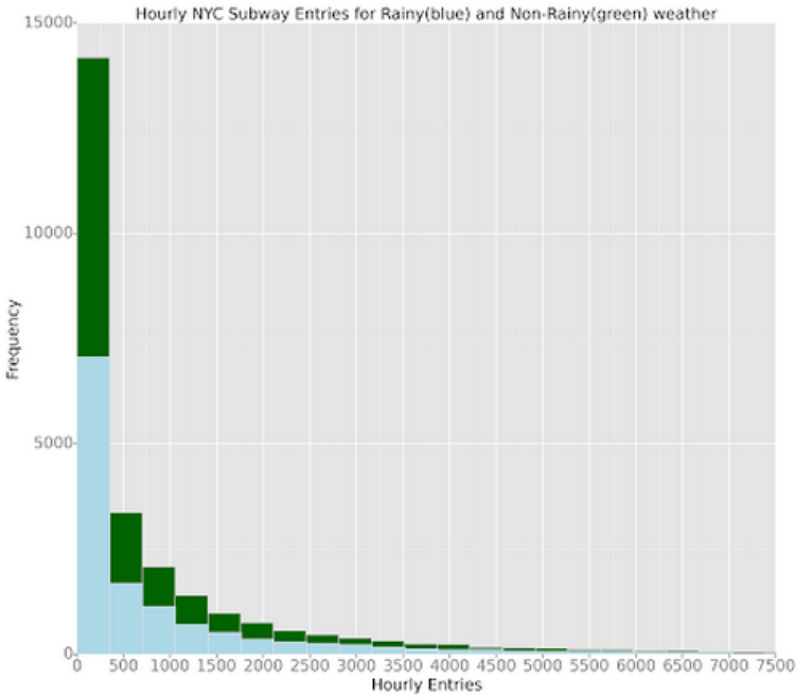
2.5: $R^2 = 0.463968815042$

2.6A: I think there may be a few other variables that can push the R^2 closer to 1, however I found this satisfactory for the model we were building.

2.6B: I think the MWU test was more applicable than this linear model. I would use MWU if I wanted to show someone a more definitive result.

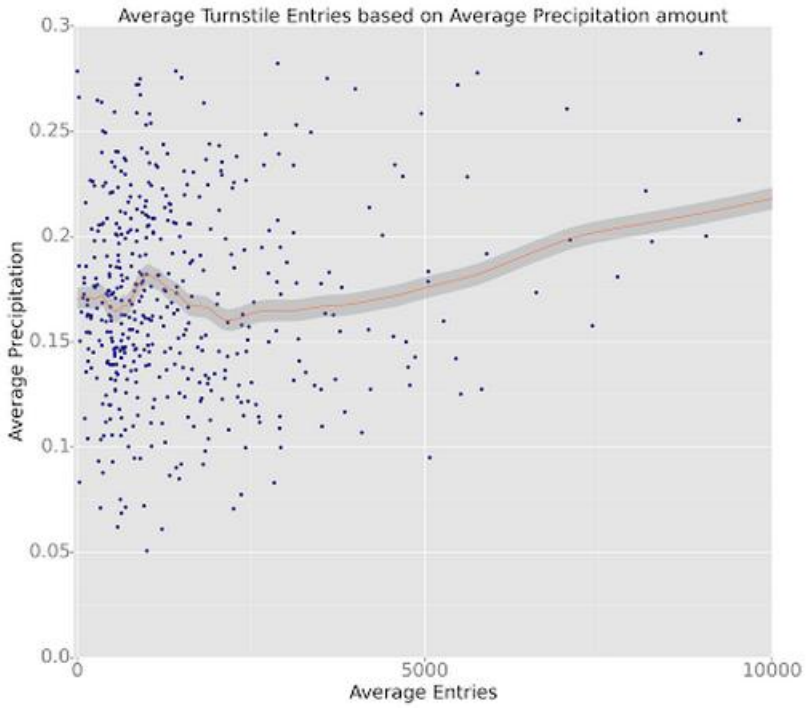
Section 3. Visualization

In []:
```
from pandas import *
from ggplot import *

def plot_weather_data(turnstile_weather):


    df_rain = turnstile_weather[turnstile_weather.rain == 1]
    df_rain = df_rain.reset_index(drop=True)

    df_no_rain = turnstile_weather[turnstile_weather.rain == 0]
    df_no_rain = df_no_rain.reset_index(drop = True)

    width = 350

    plot = ggplot(df_no_rain, aes(x = "ENTRIESn_hourly")) +\
           geom_histogram(fill = "darkgreen", binwidth = width) +\
           geom_histogram(df_rain, aes(x = "ENTRIESn_hourly"), fill = "lightblue", binwidth = width) +\
           theme(text = element_text(size=20)) +\
           scale_x_continuous(limits = (0,7500)) +\
           scale_y_continuous(limits = (0,15000)) +\
           ggtitle("Hourly NYC Subway Entries for Rainy(blue) and Non-Rainy(green) weather ") +\
           xlab("Hourly Entries") +\
           ylab("Frequency")


    return plot
```



3.1: The above histogram shows NYC subway ridership instances for rain and no rain aggregate data. Since the data is not normally distributed in this case, our above visualization does a poor job of proving our finding that ridership increases while it is raining.

In []:
```
from ggplot import *
import pandas
from pandasql import sqldf


def plot_weather_data(turnstile_weather):


    turnstile_weather = turnstile_weather.rename(columns= {'Unnamed: 0':'i'})

    q = '''
          SELECT `UNIT`, AVG(`precipi`) as avg_precip, AVG(ENTRIESn_hourly) as avg_entries
          FROM turnstile_weather
          GROUP BY `UNIT`
        '''

    data = sqldf(q,locals())

    plot = ggplot(data,aes('avg_entries','avg_precip')) +\
    geom_point(color='darkblue') +\
    theme(text = element_text(size=24)) +\
    stat_smooth(span=.25, color='#FF6600', se=True) +\
    scale_x_continuous(limits = (-50,10000)) +\
    scale_y_continuous(limits = (0,.30)) +\
    ggtitle('Average Turnstile Entries based on Average Precipitation amount') +\
    xlab("Average Entries") +\
    ylab("Average Precipitation")

    return plot
```



3.2: The above graph shows a correlation in ridership and precipitation. We can see that some of the largest amounts for turnstile entries took place when the average precipitation was higher.

Section 4. Conclusion

4.1: According to my analysis, more people ride the NYC subway when it is raining.

4.2: Looking at my Mann-Whitney U-test you can see that the mean ridership for rainy days is higher than those of non-rainy days. The P-Value also shows a 2.5% error which is very low and in my mind, can clearly show that NYC subway ridership is likely to be greater while raining.

My linear regression model didn't show as much of a correlation between rainy hours and ridership. The difference was negligible. However, that is why we tested using both methods in cases like this. I would assume this has to do with sample sizes.

Section 5. Reflection

5.1A: The biggest shortcoming of the dataset for me was the bigger picture. We truly don't know if the sample sizes/populations are skewed. Some people only ride the subway to work, a subway station may be closed for construction, or maybe there was a large event such as a Yankees or Mets baseball game, for example. There are a lot of other variables at play that we are not given in this dataset.

5.1B: The biggest shortcoming of the analysis methods that I used was probably being new to this altogether. I haven't perfected these methods and I think that with time I can squeeze a better analysis out of both methods. Other than that, I would say we should zero in on only a few turnstile units and focus on analyzing it over a longer period of time. This would be a better indicator, as our sample population may be more consistent, thus increasing the efficiency of our results.