TpT Assignment: Kaggle Bike Share Challenge

Divya Sriram

June 2017

MORE DATA, MORE QUESTIONS

1. Where is this climate data from?
2. Train-Test split

   Train = 1st – 19th

   Test = 20th – 30/31st
3. DC has weird seasons according to the data…

Spring	=	Jan - Mar
Summer	=	Apr - Jun
Fall	=	Jul - Sept
Winter	=	Oct – Dec
4. Units??

   Windspeed – mph?

   Humidity - ?

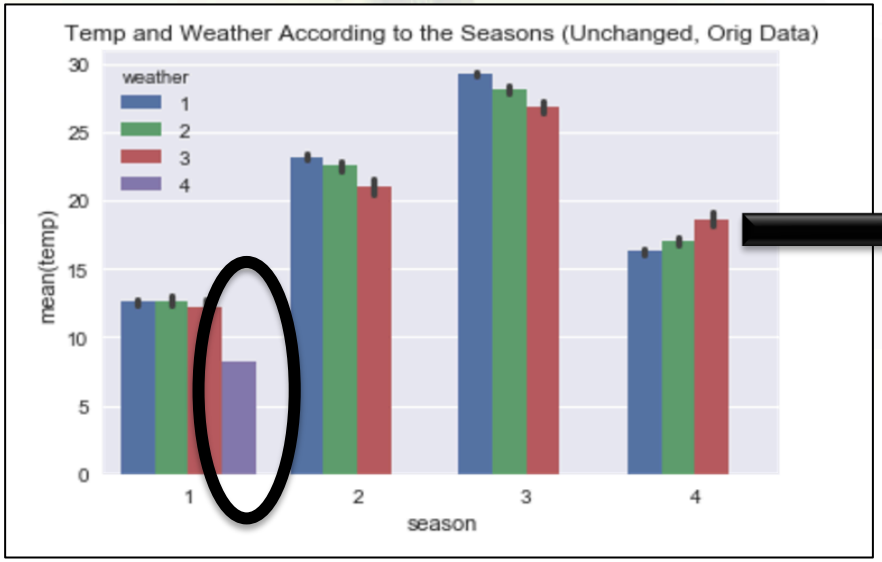
   Useful to do some sanity checks / ball park likelihoods

ASSUMPTIONS + MISSING GENERAL INFO

1. Missing rows of data !
   - Some hours of certain days are missing
   - Dense enough data, didn't impute these rows
2. First 2/3 of the month ≈ Last 1/3 of the month?
   - Do locals make up population for first 2/3 of month & visitors + tourists make up greater part of last 1/3 of month?
3. Working Population
   - Workinday or not – is a variable
   - Usefulness of the variable depends on;
     - What portion of the population has a steady job?
     - Unemployed?
4. Location of bike shares?
   - Certain locations get greater traffic for rentals?
5. Bike event dates?
   - Bike event → greater probability for higher # rentals

CHANGING SEASONS

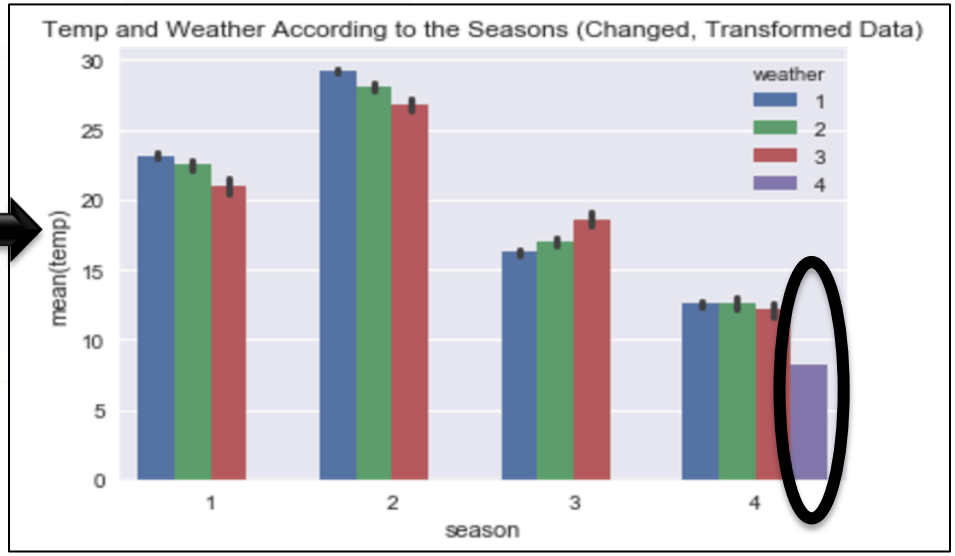
UNCHANGED SEASONS



1. Incorrect season/month grouping
2. Temperature Trend – weird!

coldest → warmer → hottest → cool down
(spring) (summer) (fall) (winter)
(1) (2) (3) (4)

CHANGED SEASONS



1. Sensible season/month grouping
2. Temperature Trend – normal!

Warmer → hottest → cool down→ coldest
(spring) (summer) (fall) (winter)
(1) (2) (3) (4)

SANITY CHECKS FOR DATA SET

- Holidays , Weekends , Working Days
  - If it's a holiday:

    → it can't be a working day ✔
  - If it's not a holiday & not the weekend

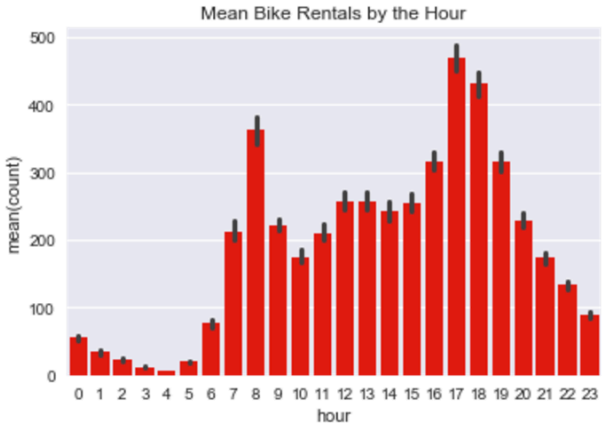
    → it has to be a working day ✔

- Are the bike renters an anomaly?
  - Most people prefer nice weather to bike
  - Dataset shows people prefer nice weather ✔

INTERESTING GLIMPSES

- Max Rentals in 1 Hour = 977!
  - Sept 9, 2012 (6:00pm – 7:00pm)
  - What was happening Sept 9, 2012?
  - DC Bike Party (meet up time = 7:30pm)

- Do people bike at odd hours?
  - You can rent for 24 hrs
  - Forgot to return?
  - Just keeping overnight not riding



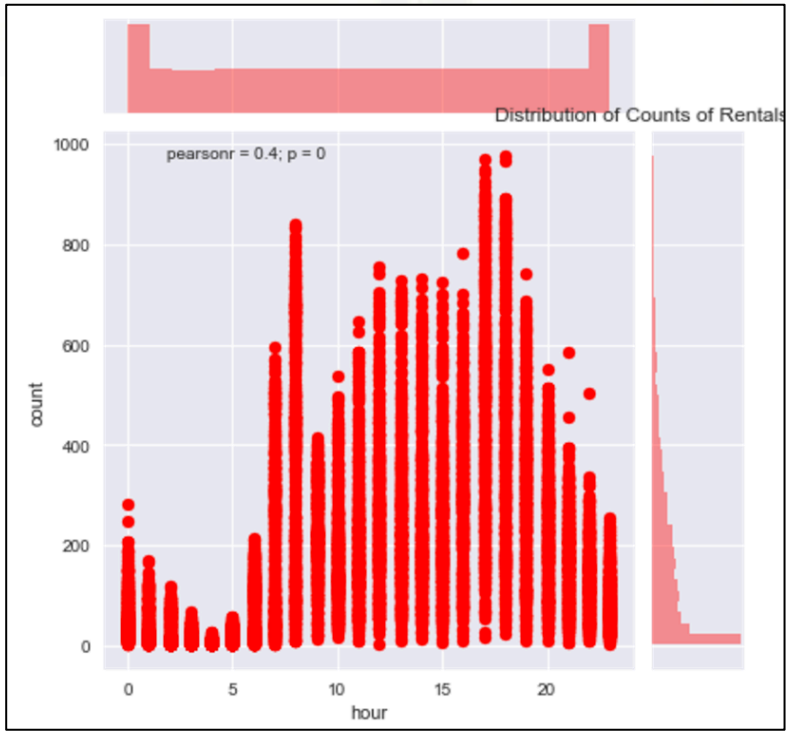
RELATIONSHIP BETWEEN VARIABLES

AND

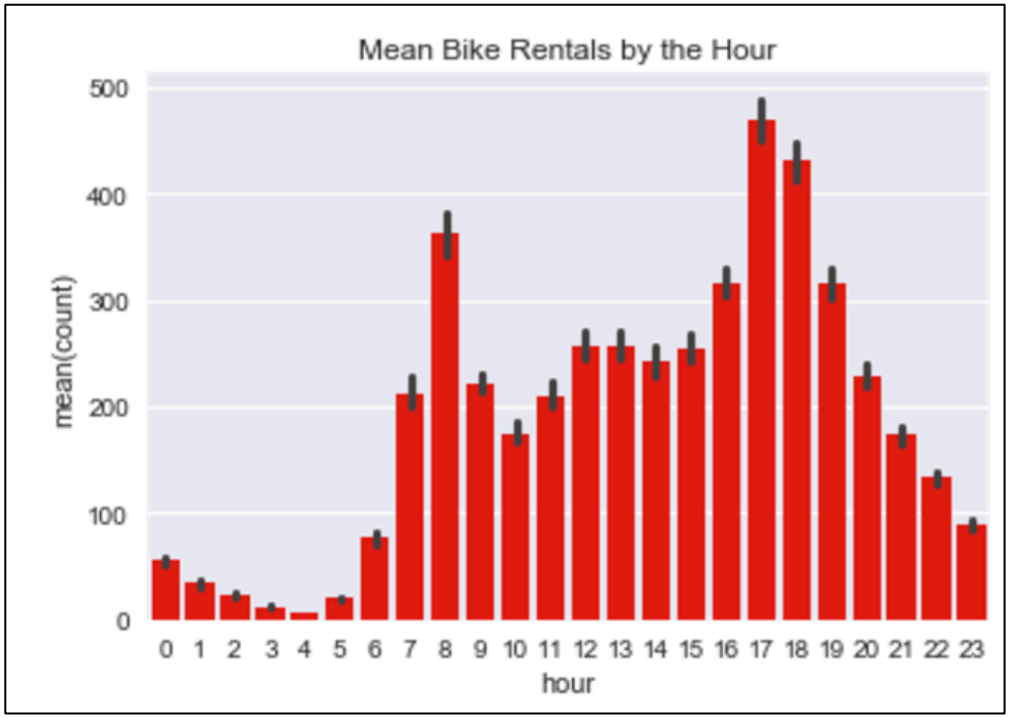
OUTCOME VARIABLE (COUNT)

HOUR vs COUNT

Distribution of Counts of Rentals by the Hour

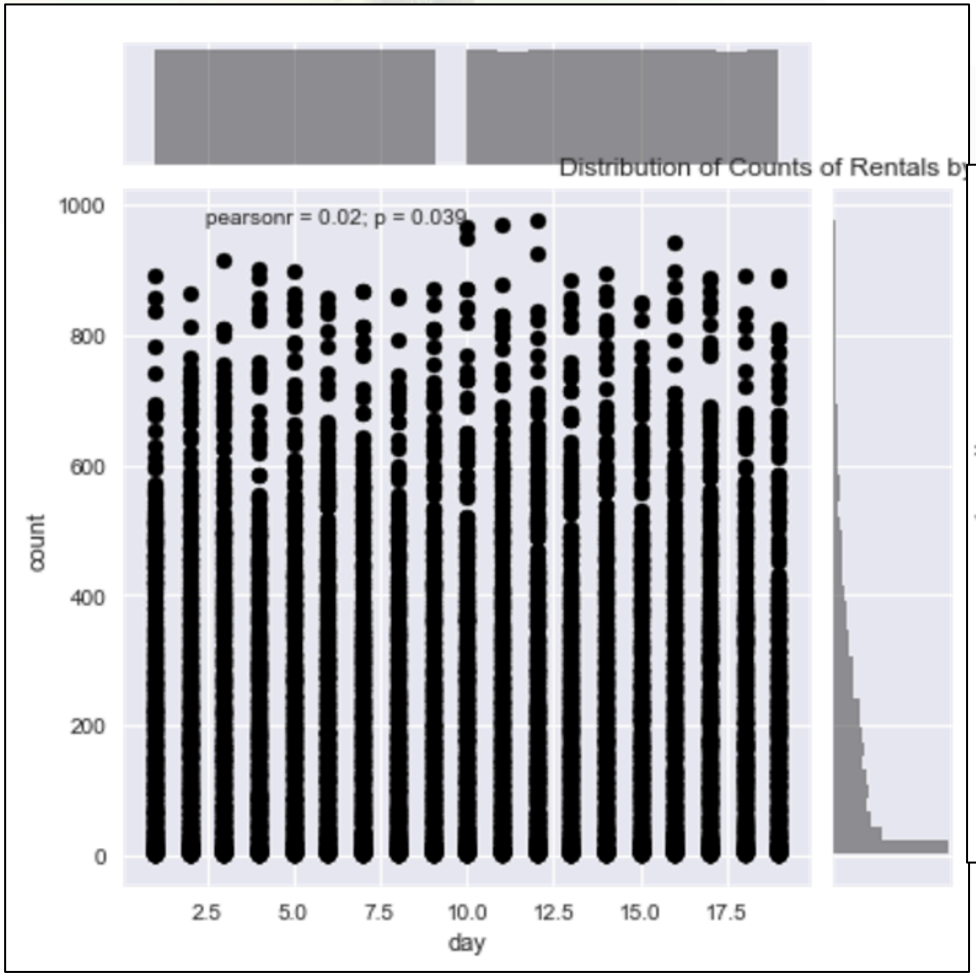


Mean Bike Rentals by the Hour

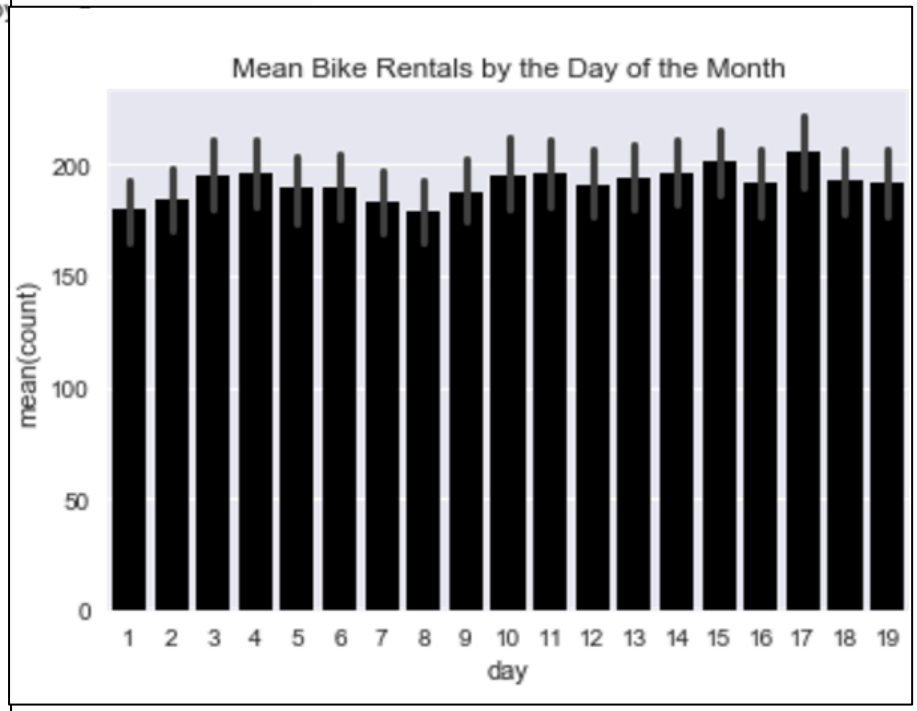


DAY OF MONTH vs COUNT

Distribution of Counts of Rentals by the Day of the Month

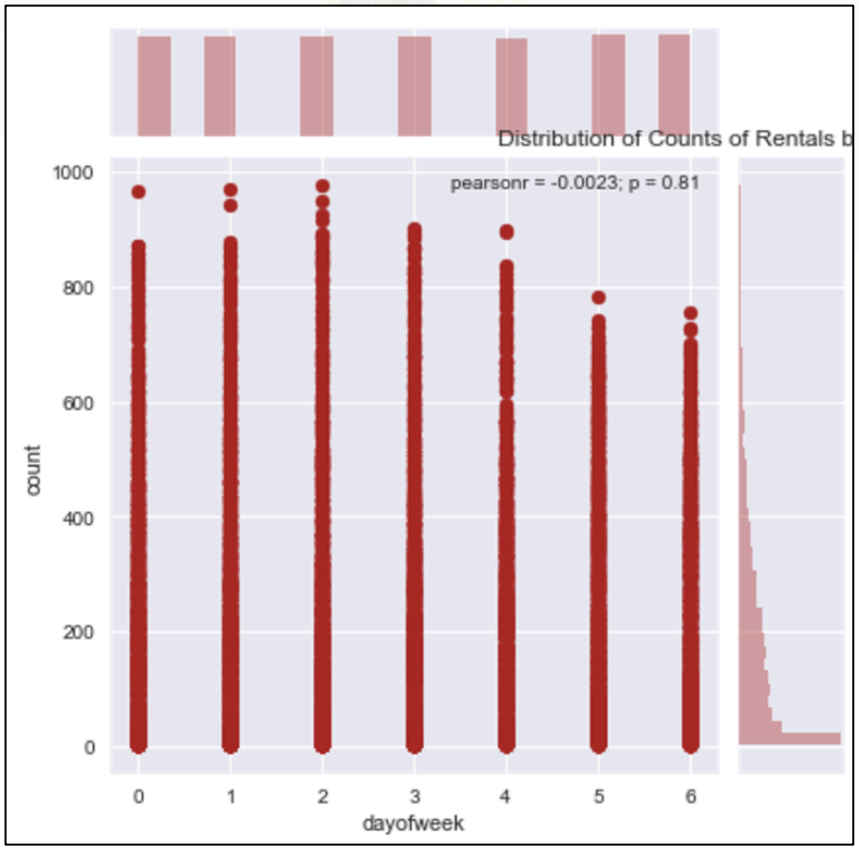


Mean Bike Rentals by the Day of the Month

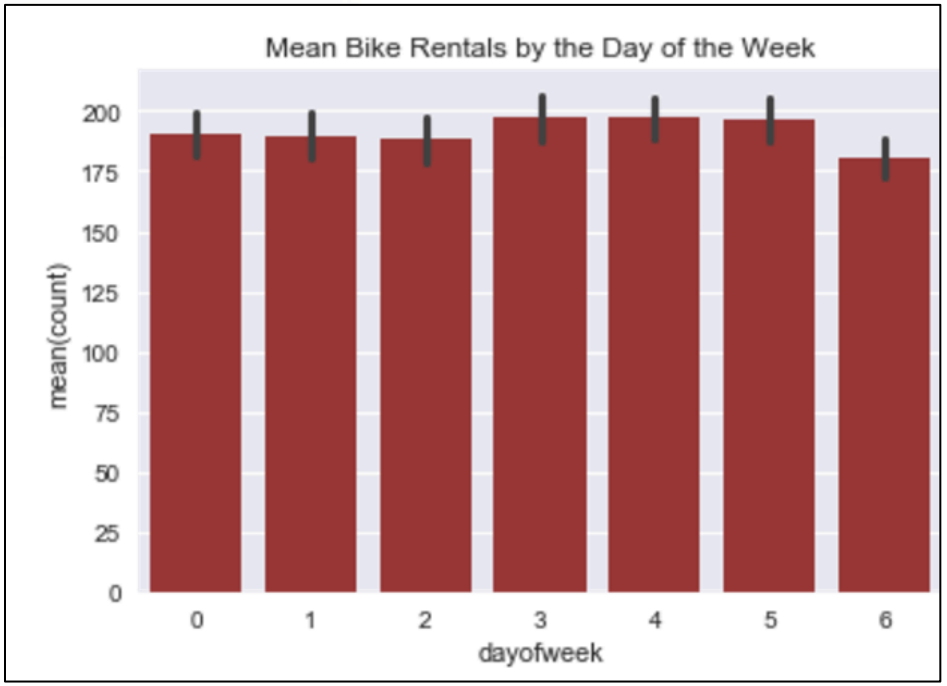


DAY OF WEEK vs COUNT

Distribution of Counts of Rentals by the Day of the Week

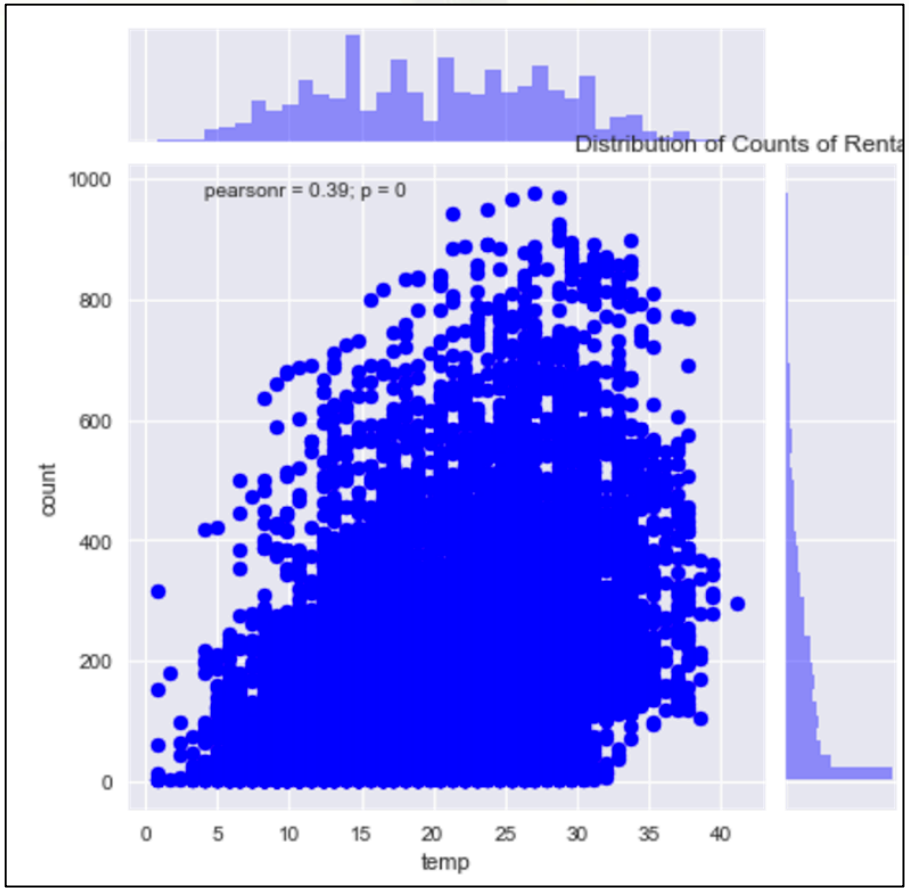


Mean Bike Rentals by the Day of the Week

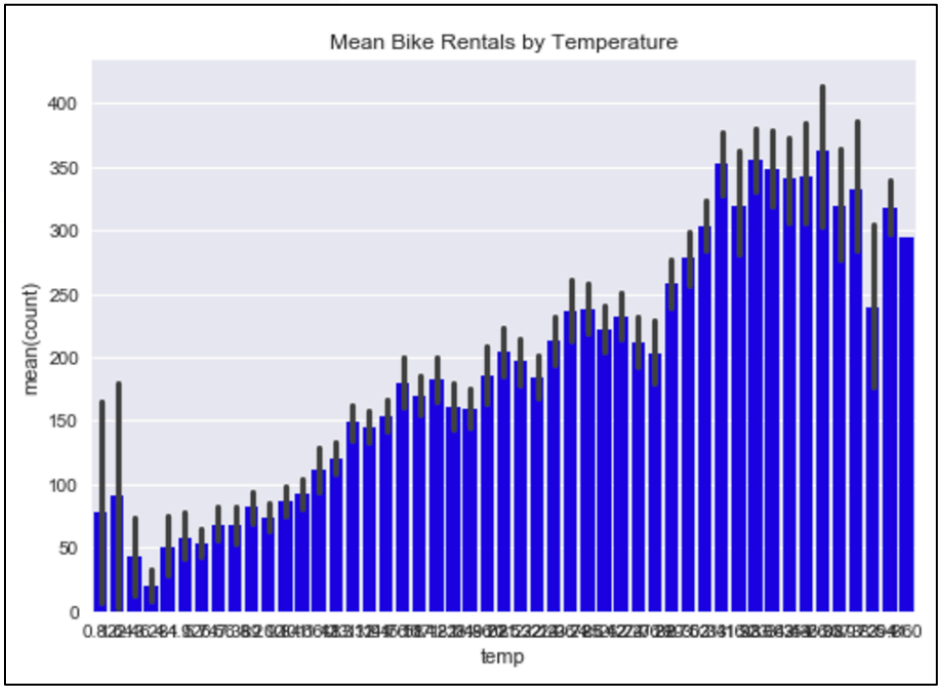


TEMPERATURE vs COUNT

Distribution of Counts of Rentals by Temperature

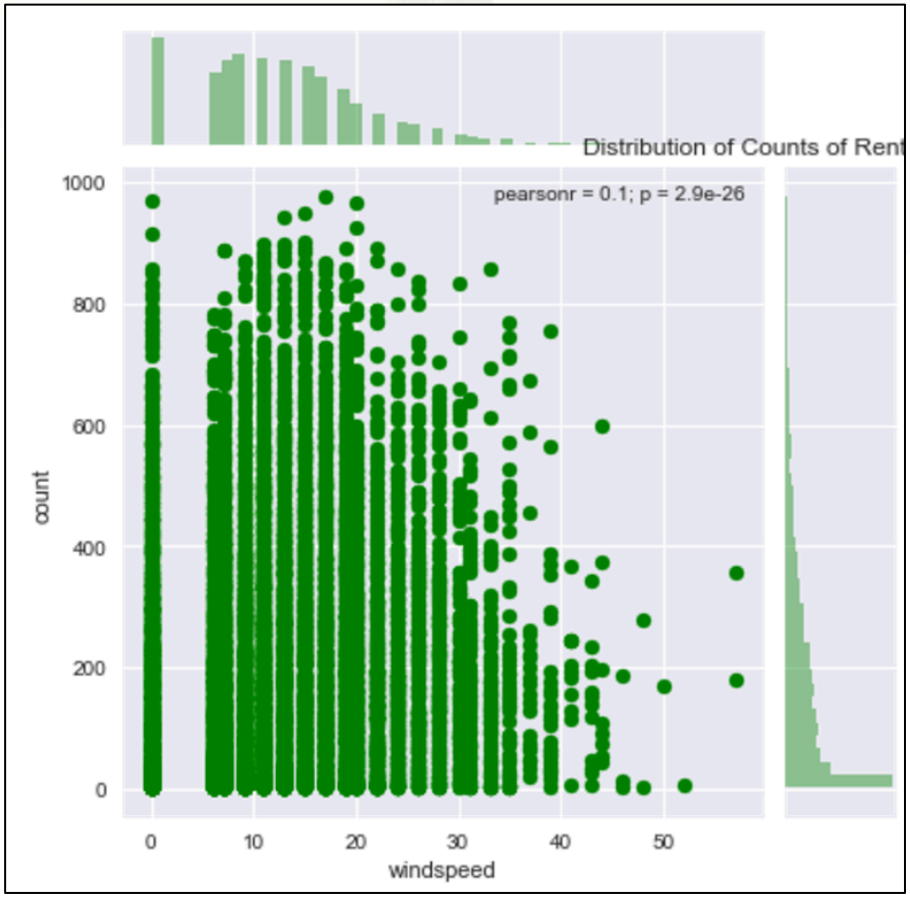


Mean Bike Rentals by Temperature

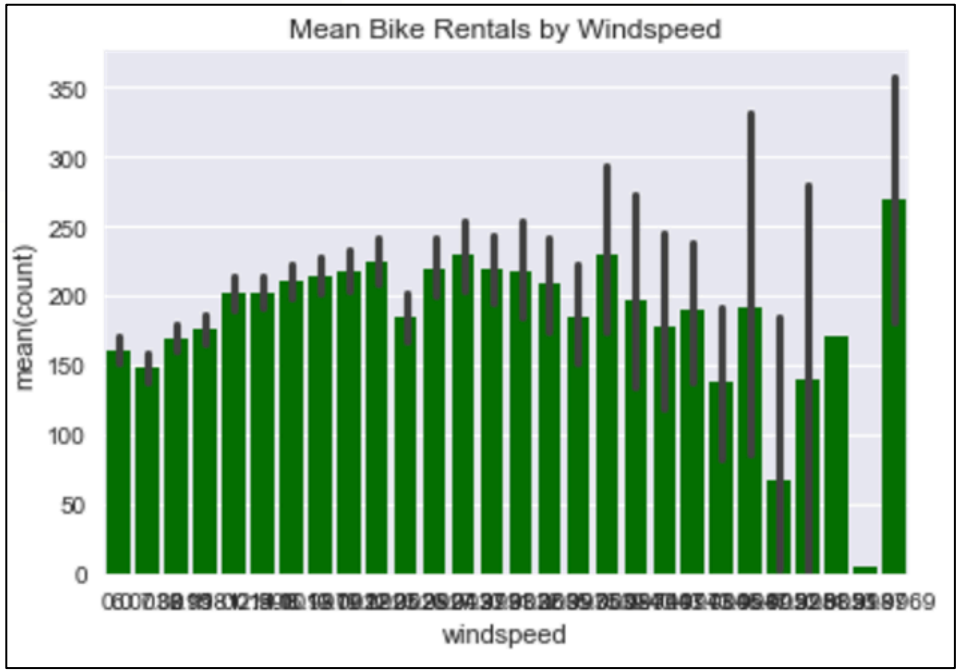


WINDSPEED vs COUNT

Distribution of Counts of Rentals by Windspeed

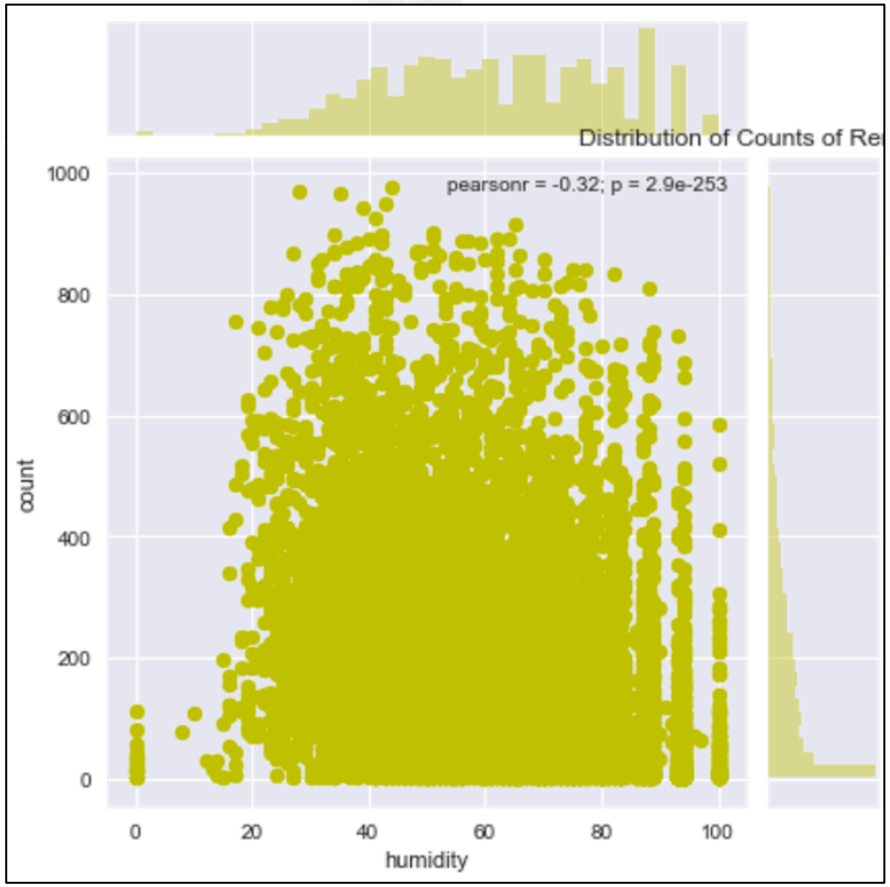


Mean Bike Rentals by Windspeed

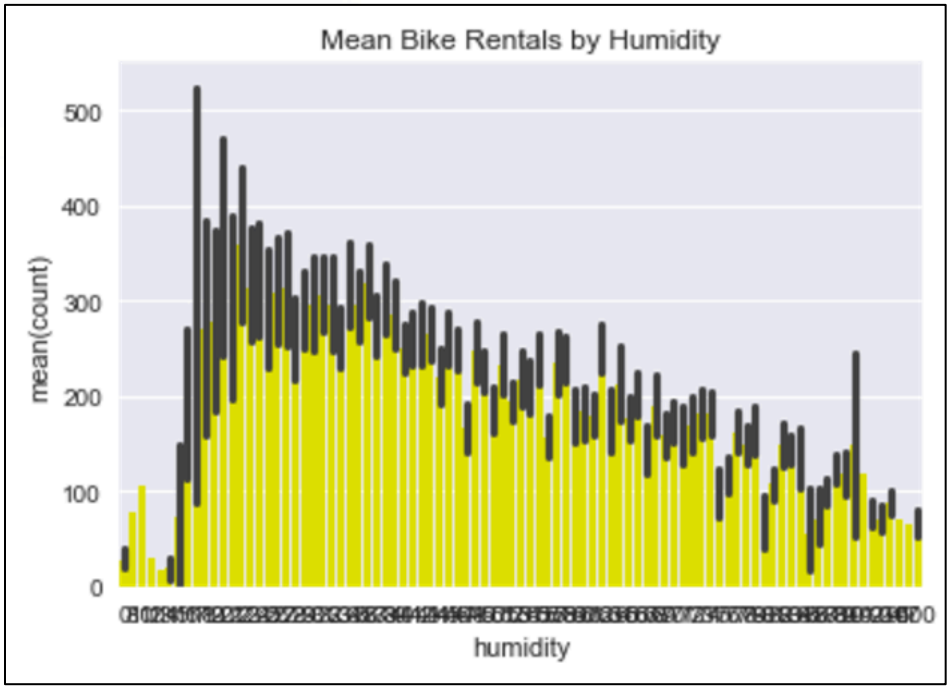


HUMIDITY vs COUNT

Distribution of Counts of Rentals by Humidity

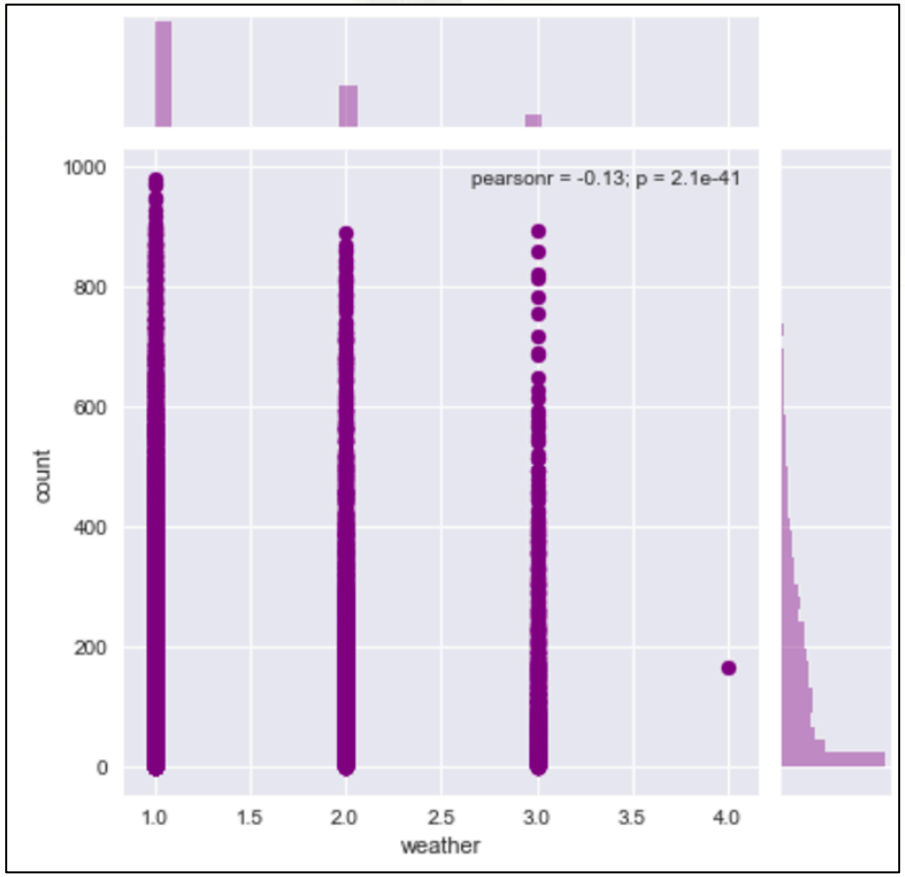


Mean Bike Rentals by Humidity

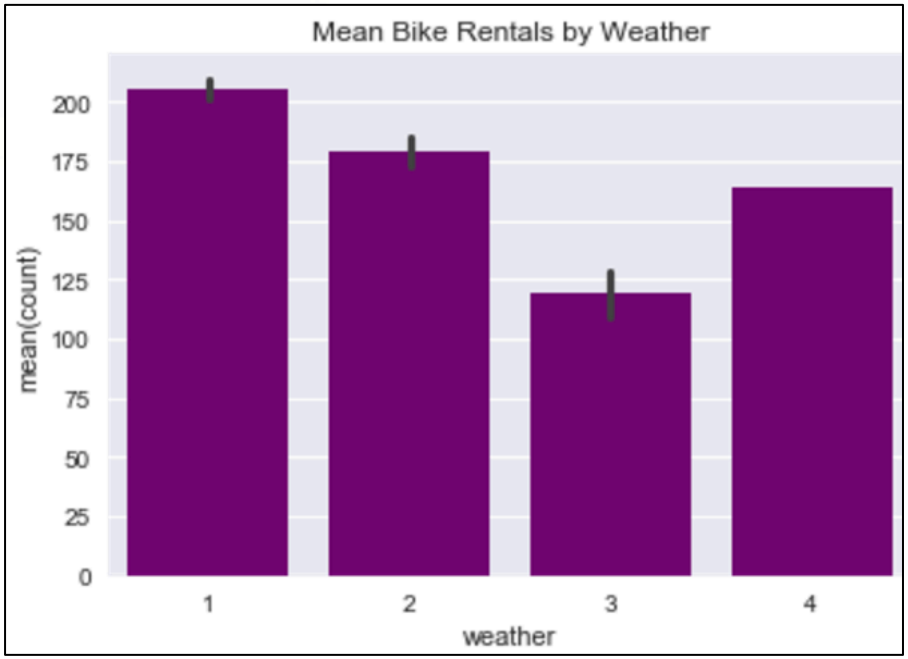


WEATHER vs COUNT

Distribution of Counts of Rentals by Weather

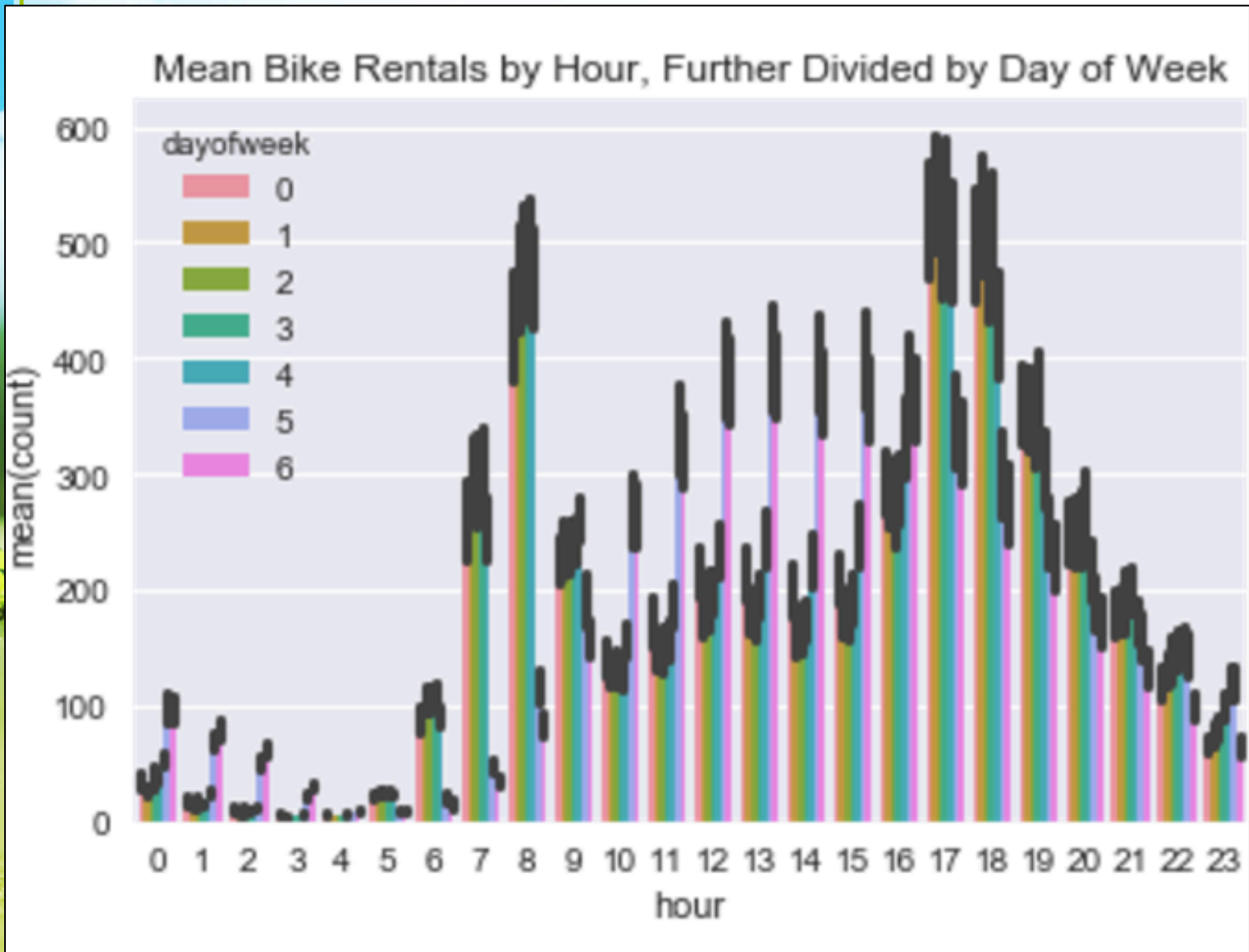


Mean Bike Rentals by Weather

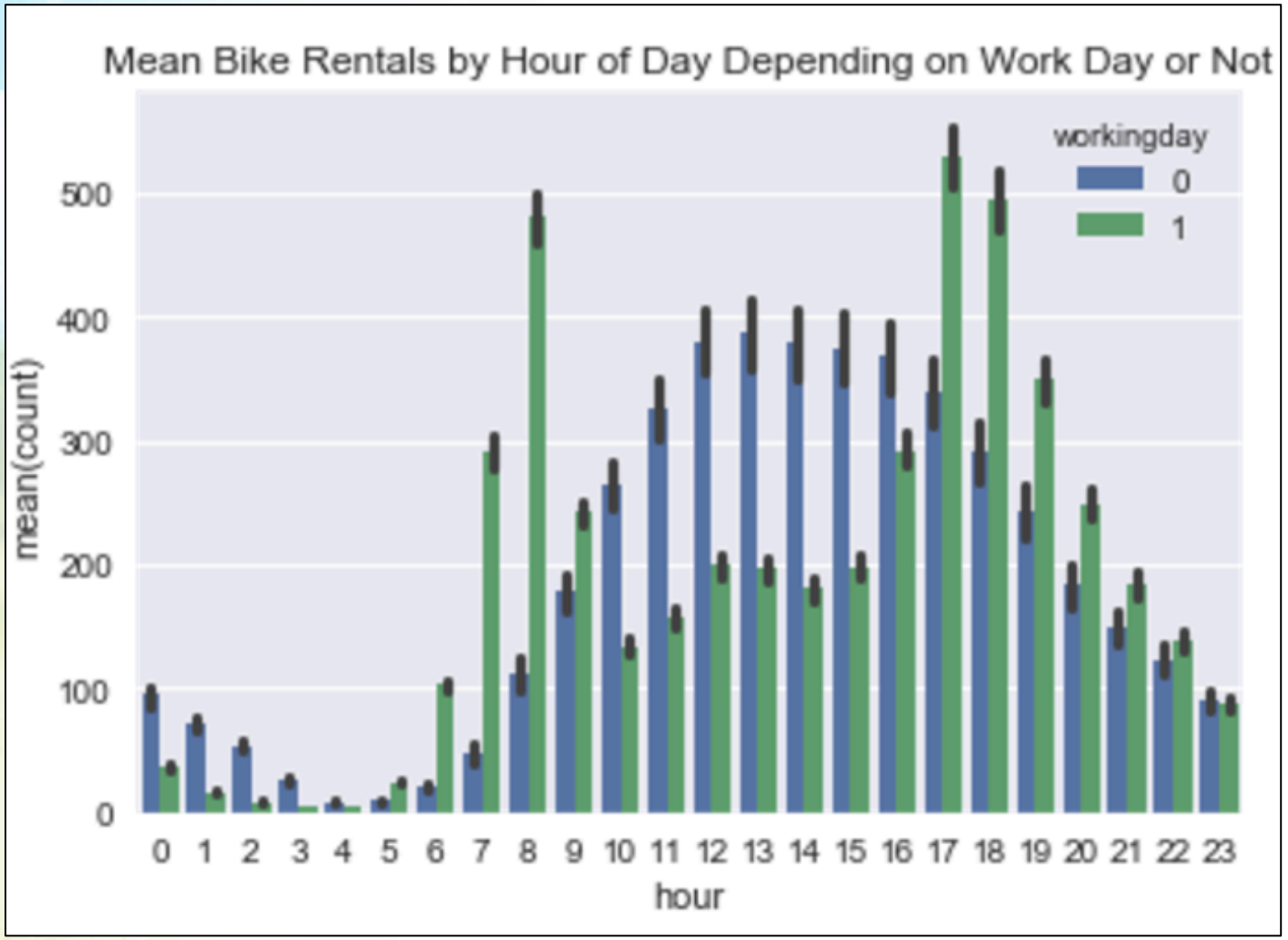


OBSERVING VARIABLES TOGETHER

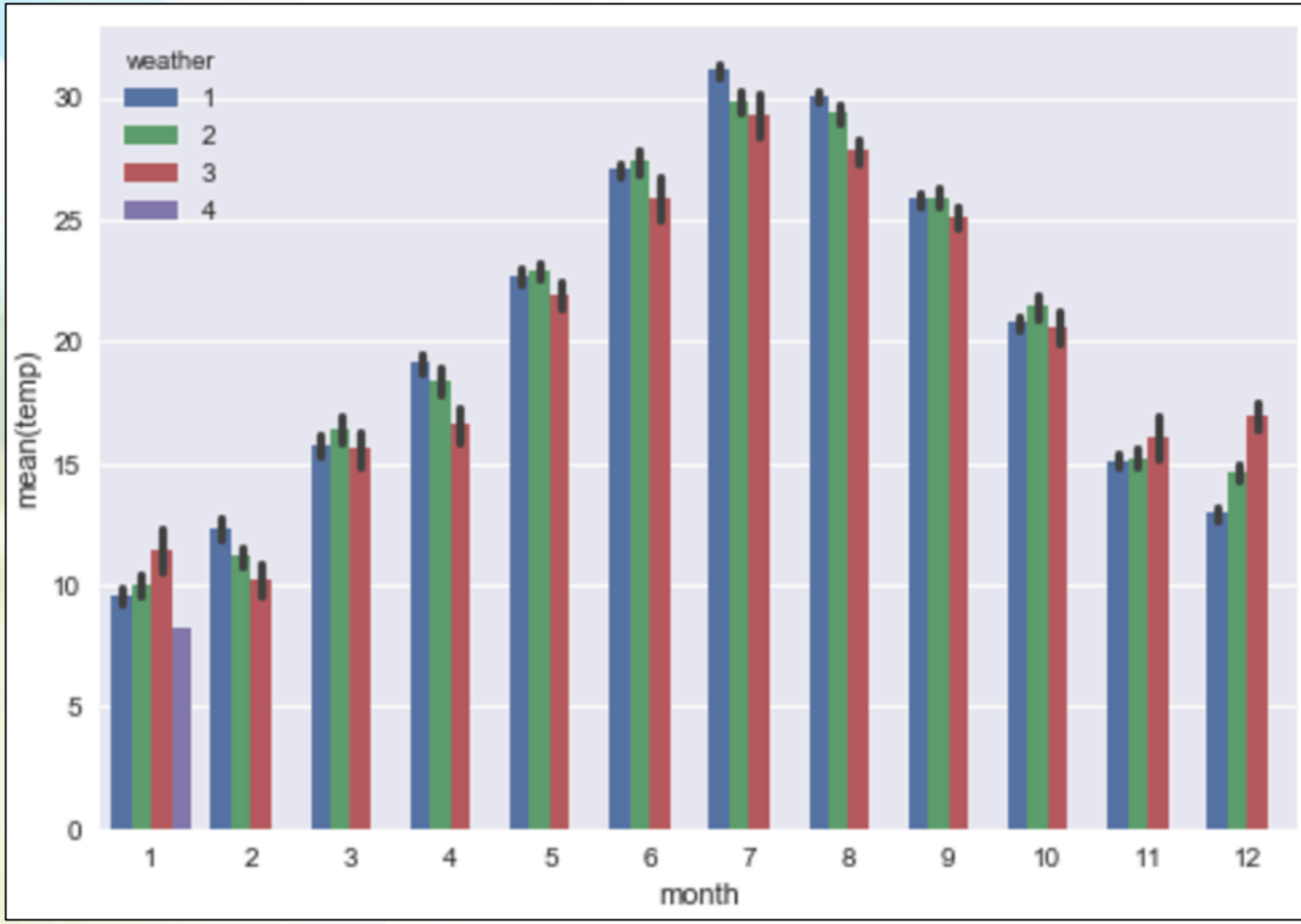
Mean Bike Rentals by Hour of Day, Further Divided by Day of Week



Mean Bike Rentals by Hour of Day Depending on Work Day or Not



Mean Temperatures by Month, Further Divided by Weather



VARIABLES I LEFT OUT FOR NOW

- atemp
- casual
- registered

PREDICTION MODEL(S)

WHAT CRYSTAL BALL SHOULD I USE TO PREDICT YOUR FUTURE:

- RPART
  - Highest error (0.90215)

- PARTY (ctree)
  - Lower error (0.67175)

- RANDOM FORESTS
  - Lowest error (0.60693)

** changed vs unchanged seasons (random forests) :
(0.60693) (0.59960)

BIASES IN THE DATA

HOW DO THEY IMPACT WHAT WE DO WITH OUR MODEL FOR THE FUTURE?

1. Excluding groups of people in the data
    - Disabled people
    - People who can't afford to pay the fee
2. Lack of location information
    - Does it truly represent ALL of DC?
3. Assuming other factors stay the same?
    - Fee to rent
    - Infrastructure exists

NEXT STEPS

- Try multiple linear regression?
- Try other ML model with python packages
- Better feature engineering!
- Build model to predict + fill in 1000 records with 0mph windspeed

THANK YOU

KAGGLE USERNAME: “ divyasriram ”

Divya Sriram
Master of Information and Data Science (MIDS 2018)
divyasriram@berkeley.edu