DATA 606 Data Project Proposal

Donald Butler

11/07/2021

Data Preparation

```
library(tidyverse)

bike_data <- read_csv('https://raw.githubusercontent.com/dab31415/DATA606/main/Project_1/day.csv')
```

Research question

Is temperature an accurate predictor for bicycle ride sharing?

Cases

Each observation is for a single day and contains the number of bicycles rented by casual and registered users in Washington DC. The observation also contains weather statistics for temperature, humidity, wind speed, and precipitation.

Data collection

Bike Sharing data was collected through https://capitalbikeshare.com which provides bicycle rentals in Washington, DC. The researched collected two years of bike sharing data then added corresponding weather and seasonal information.

Type of study

This is an observational study.

Data Source

The data was obtained at https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset, and was originally published in 2013. (Fanaee-T and Gama 2013)

Dependent Variable

The dependent variable is the number of bicycles that were rented each day.

Independent Variable

The independent variables are temperature and weather type.

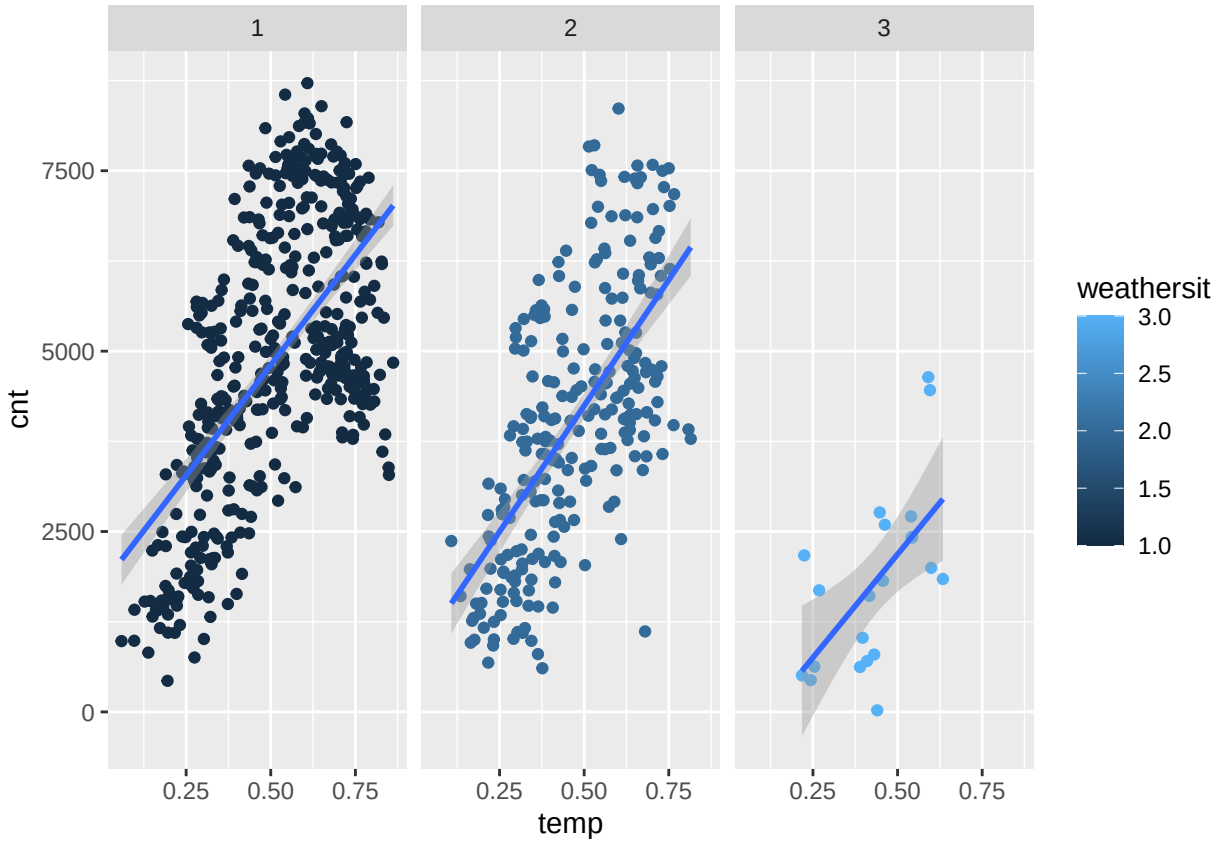
Relevant summary statistics

```
summary(bike_data)
```

```
##     instant          dteday              season           yr
##  Min.   :  1.0   Min.   :2011-01-01   Min.   :1.000   Min.   :0.0000
##  1st Qu.:183.5   1st Qu.:2011-07-02   1st Qu.:2.000   1st Qu.:0.0000
##  Median :366.0   Median :2012-01-01   Median :3.000   Median :1.0000
##  Mean   :366.0   Mean   :2012-01-01   Mean   :2.497   Mean   :0.5007
##  3rd Qu.:548.5   3rd Qu.:2012-07-01   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :731.0   Max.   :2012-12-31   Max.   :4.000   Max.   :1.0000
##       mnth          holiday           weekday        workingday
##  Min.   : 1.00   Min.   :0.00000   Min.   :0.000   Min.   :0.000
##  1st Qu.: 4.00   1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.000
##  Median : 7.00   Median :0.00000   Median :3.000   Median :1.000
##  Mean   : 6.52   Mean   :0.02873   Mean   :2.997   Mean   :0.684
##  3rd Qu.:10.00   3rd Qu.:0.00000   3rd Qu.:5.000   3rd Qu.:1.000
##  Max.   :12.00   Max.   :1.00000   Max.   :6.000   Max.   :1.000
##    weathersit         temp             atemp              hum
##  Min.   :1.000   Min.   :0.05913   Min.   :0.07907   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.33708   1st Qu.:0.33784   1st Qu.:0.5200
##  Median :1.000   Median :0.49833   Median :0.48673   Median :0.6267
##  Mean   :1.395   Mean   :0.49538   Mean   :0.47435   Mean   :0.6279
##  3rd Qu.:2.000   3rd Qu.:0.65542   3rd Qu.:0.60860   3rd Qu.:0.7302
##  Max.   :3.000   Max.   :0.86167   Max.   :0.84090   Max.   :0.9725
##    windspeed          casual         registered        cnt
##  Min.   :0.02239   Min.   :   2.0   Min.   :  20   Min.   :  22
##  1st Qu.:0.13495   1st Qu.: 315.5   1st Qu.:2497   1st Qu.:3152
##  Median :0.18097   Median : 713.0   Median :3662   Median :4548
##  Mean   :0.19049   Mean   : 848.2   Mean   :3656   Mean   :4504
##  3rd Qu.:0.23321   3rd Qu.:1096.0   3rd Qu.:4776   3rd Qu.:5956
##  Max.   :0.50746   Max.   :3410.0   Max.   :6946   Max.   :8714
```

```
bike_data %>%
  ggplot(aes(temp,cnt)) +
  geom_point(aes(color = weathersit)) +
  geom_smooth(method = lm) +
  facet_wrap(~ weathersit)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Fanaee-T, Hadi, and Joao Gama. 2013. "Event Labeling Combining Ensemble Detectors and Background Knowledge." *Progress in Artificial Intelligence*, 1–15. https://doi.org/10.1007/s13748-013-0040-3.