Analysis of Variance

(ANOVA)

Prof. Maria Tackett

02.03.20

STA 210

Click for PDF of slides

Announcements - UPDATE!!!

- Lab 03 - **due Tuesday at 11:59p**
- Reading

rstudio::conf 2020 Tweets

- As ML becomes ubiquitous in the industry, it is critical that we discover ways to explain the under-the-hood workings of ML in human terms (for non-technical users).

- "Presenters at rstudio::conf design and develop curriculum to democratize data science pedagogy beyond elite universities and highly educated people, aiming to promote data literacy and economic empowerment for many.""

- "Jenny Bryan says that the smaller the 'haystack' is the easier it is to find the error. ie reduce amount of code that error could be located."

STA 210

rstudio::conf 2020 Tweets

- "No matter how impactful your results are, your data/message will only be as good as the visualization you create. Take the time and effort to make sure your story is conveyed clearly and ~beautifully~. Graph aesthetics are more important than you think!"

- "@SharlaGelfand Likewise, I often can't decipher my notes. As an R beginner, I thought this personality trait made me unfit to use R, but your talk has convinced otherwise. I'm inspired to implement R into my daily life from now on!"

Today's Agenda

- Comparing group means using Analysis of Variance (ANOVA)

Capital Bike Share

The Capital Bike Share is a bike share program in Washington D.C. where customers can rent a bike for a small fee, ride it around the city, and return it to a station located near their destination

Bike riding is often correlated with environmental conditions, so we are going to analyze the relationship between season (**`season`**) and the number of daily bike rentals (**`count`**)

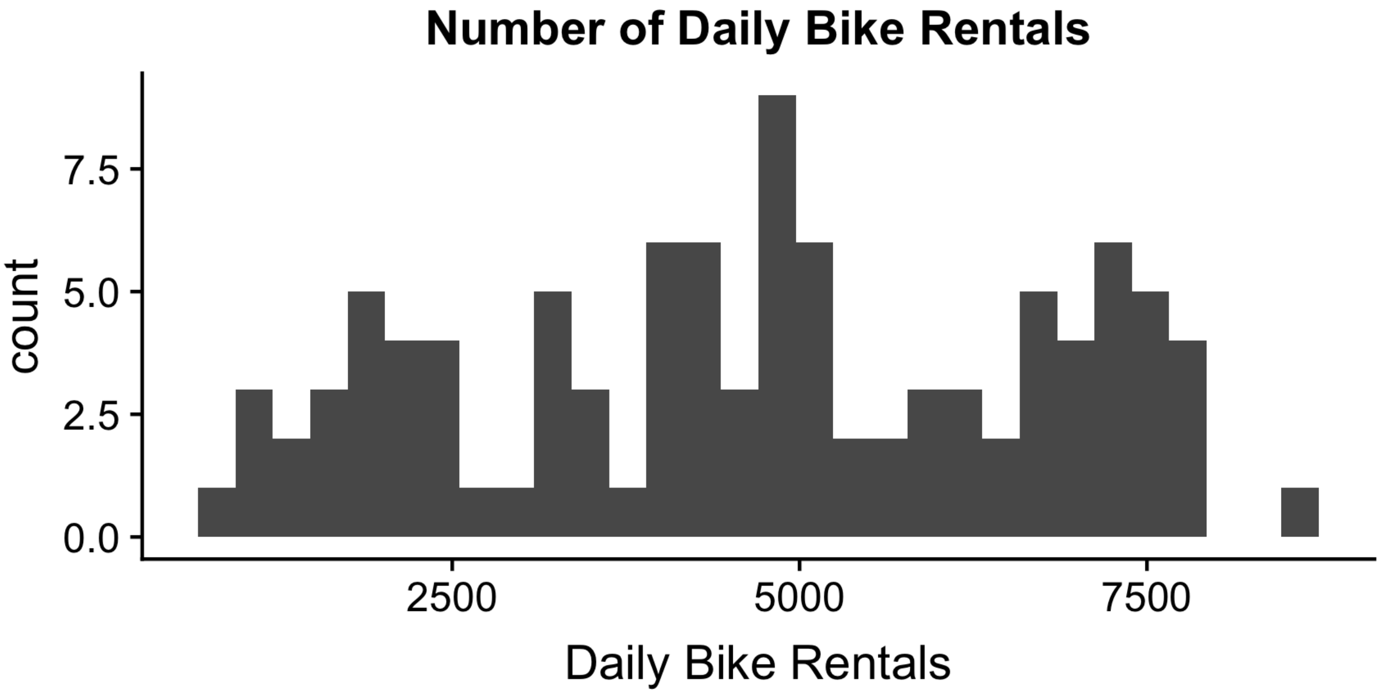
Capital Bike Share

Our dataset contains the number of bikes rented and other information for **100 randomly selected days** in 2011 and 2012

```r
bikeshare <- read_csv("data/bikeshare-sample.csv")
glimpse(bikeshare)
```

```
## Observations: 100
## Variables: 16
## $ instant    <dbl> 649, 394, 125, 373, 101, 334, 308, 664, 476, 82, 651,…
## $ dteday     <date> 2012-10-10, 2012-01-29, 2011-05-05, 2012-01-08, 2011…
## $ season     <chr> "Fall", "Winter", "Spring", "Winter", "Spring", "Fall…
## $ yr         <dbl> 1, 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1,…
## $ mnth       <dbl> 10, 1, 5, 1, 4, 11, 11, 10, 4, 3, 10, 6, 3, 8, 6, 7, …
## $ holiday    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ weekday    <dbl> 3, 0, 4, 0, 1, 3, 5, 4, 5, 3, 5, 5, 3, 1, 3, 1, 2, 3,…
## $ workingday <dbl> 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
## $ weathersit <dbl> 1, 1, 1, 1, 2, 1, 2, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1,…
## $ temp       <dbl> 0.514167, 0.282500, 0.459167, 0.337500, 0.595652, 0.3…
## $ atemp      <dbl> 0.503142, 0.272721, 0.441917, 0.340258, 0.565217, 0.3…
## $ hum        <dbl> 0.630833, 0.311250, 0.444167, 0.465000, 0.716956, 0.6…
## $ windspeed  <dbl> 0.1878210, 0.2400500, 0.2953920, 0.1915420, 0.3244740…
## $ casual     <dbl> 780, 558, 614, 599, 855, 188, 470, 875, 1340, 203, 10…
```

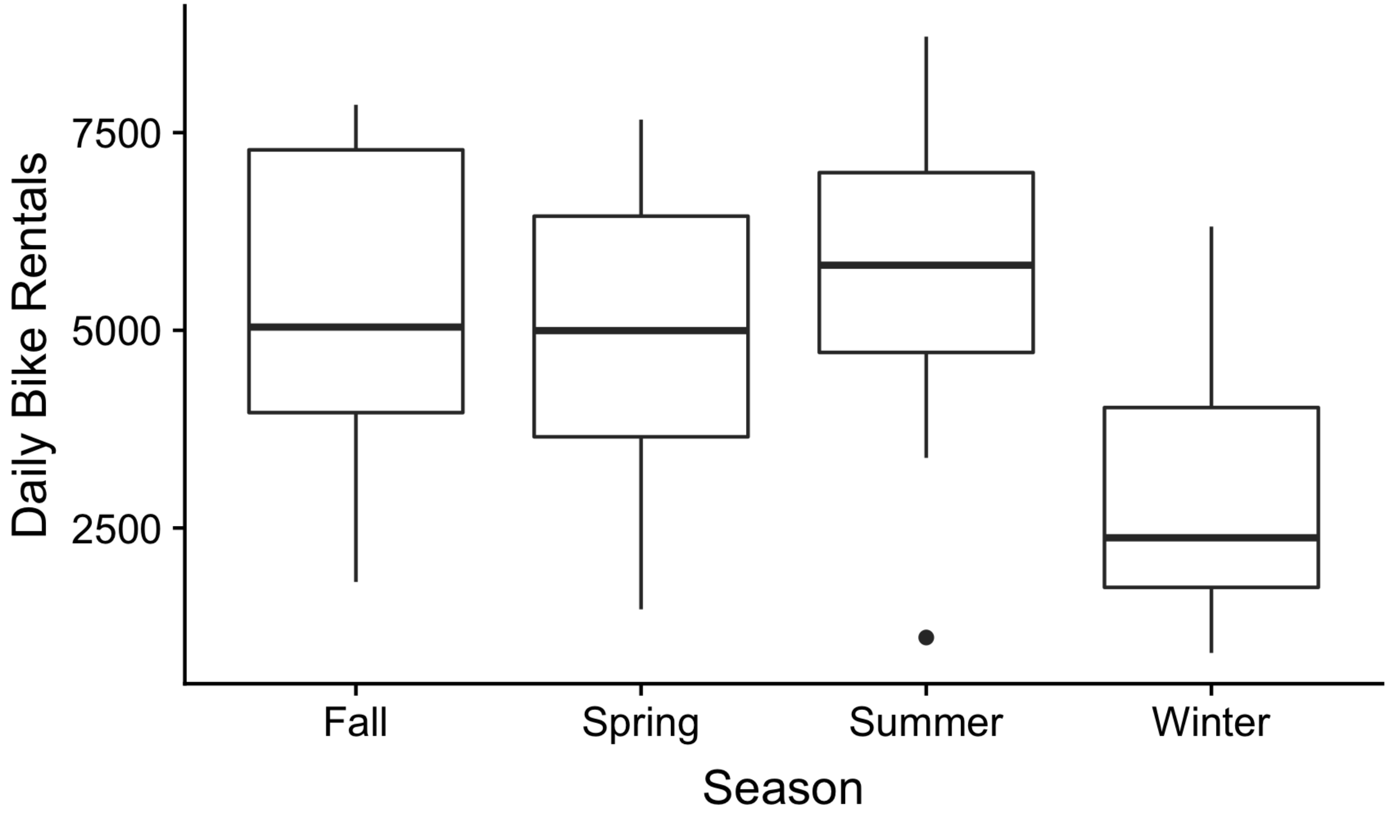
Bike rentals



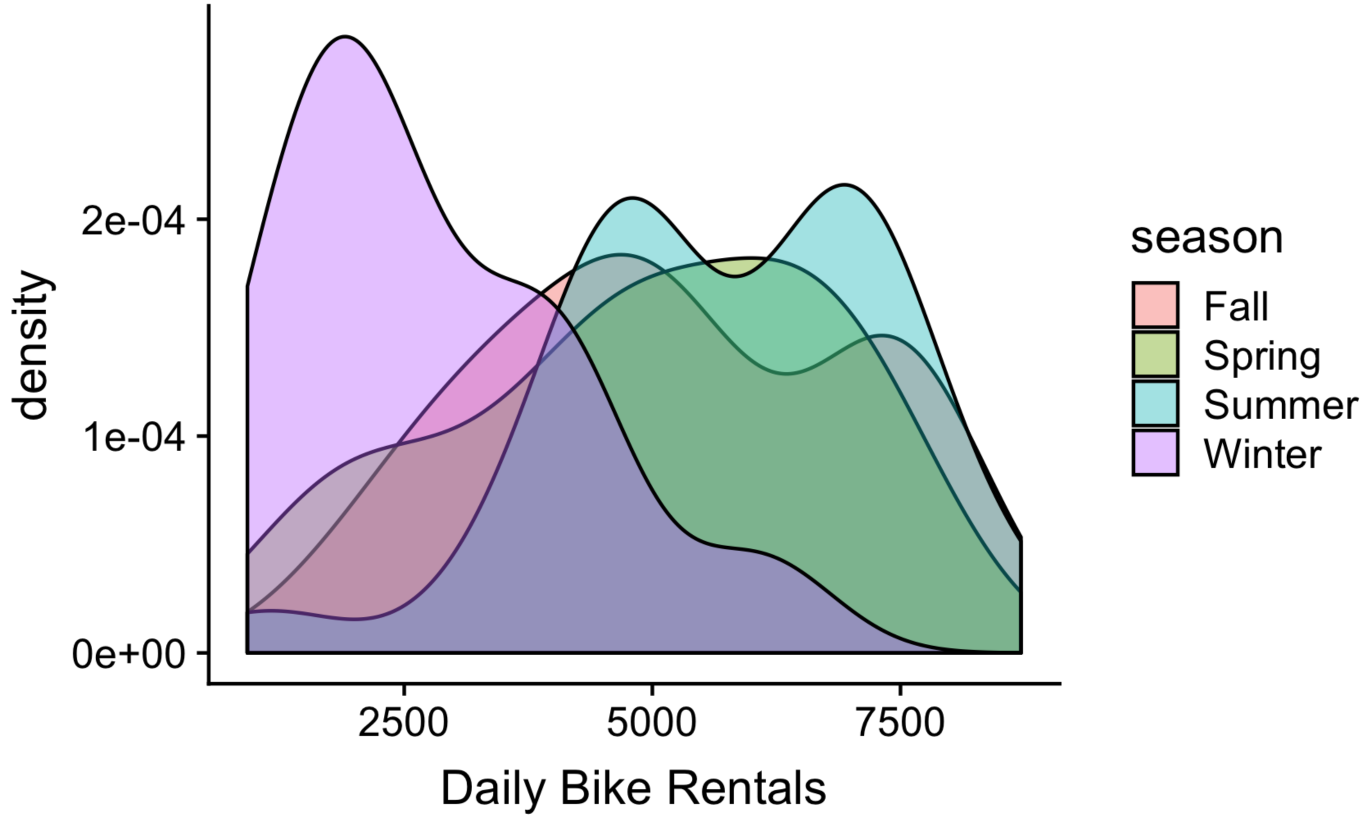

```
## # A tibble: 1 x 3
##       n  mean    sd
##   <int> <dbl> <dbl>
## 1   100 4672. 2040.
```

Question: Is there a statistically significant difference in the mean number of bikes rented in each season?

Bike rentals by season



Bike rentals by season



Bike rentals by season

```
## # A tibble: 4 x 4
##   season      n  mean    sd
##   <chr>   <int> <dbl> <dbl>
## 1 Fall       25 5180. 1848.
## 2 Spring     23 4924. 1889.
## 3 Summer     27 5739. 1662.
## 4 Winter     25 2779. 1465.
```

So far, we have used a **quantitative** predictor variable to understand the variation in a quantitative response variable.

Now, we will use a **categorical (qualitative)** predictor variable to understand the variation in a quantitative response variable.

Let's fit a model

```
bike_model <- lm(count ~ season, data = bikeshare)
tidy(bike_model, conf.int = TRUE) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	5180.200	343.843	15.066	0.000	4497.677	5862.723
seasonSpring	-256.591	496.726	-0.517	0.607	-1242.585	729.402
seasonSummer	558.911	477.178	1.171	0.244	-388.279	1506.101
seasonWinter	-2400.760	486.267	-4.937	0.000	-3365.993	-1435.527

STA 210

In-class exercise

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	5180.200	343.843	15.066	0.000	4497.677	5862.723
seasonSpring	-256.591	496.726	-0.517	0.607	-1242.585	729.402
seasonSummer	558.911	477.178	1.171	0.244	-388.279	1506.101
seasonWinter	-2400.760	486.267	-4.937	0.000	-3365.993	-1435.527

- Go to http://bit.ly/sta210-sp20-bike and use the model to answer the questions
- Use **NetId@duke.edu** for your email address.
- You are welcome (and encouraged!) to discuss these questions with 1 - 2 people around you, but **each person** must submit a response.

04:00

STA 210

How much variation is explained by our model?

Question: What proportion of the variation in number of daily bike rentals is explained by season?

```
rsquare(bike_model, bikeshare)
```

```
## [1] 0.3112098
```

About 31.12% of the variation in the number of daily bike rentals is explained by the season.

> How do we calculate this value?

STA 210

Analysis of Variance (ANOVA)

$$\hat{\text{count}} = 5180.2 - 256.591 \text{ Spring} + 558.911 \text{ Summer} - 2400.760 \text{ Winter}$$

Analysis of Variance (ANOVA) uses a single hypothesis test to check whether the means across many groups are equal*

```
anova(bike_model) %>%
  kable(format = "markdown", digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
season	3	128202929	42734310	14.458	0
Residuals	96	283747246	2955700	NA	NA

** OpenIntro Statistics* pg. 285

STA 210

Analysis of Variance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
season	3	128202929	42734310	14.458	0
Residuals	96	283747246	2955700	NA	NA

Notation

- K is number of mutually exclusive groups. We index the groups as $i = 1, \dots, K$.
- n_i is number of observations in group i
- $n = n_1 + n_2 + \cdots + n_K$ is the total number of observations in the data
- y_{ij} is the j^{th} observation in group i, for all i, j
- μ_i is the population mean for group i, for $i = 1, \dots, K$

STA 210

Motivating ANOVA

$$y_{ij} = \mu_i + \epsilon_{ij}$$

Assumption: ϵ_{ij} follows a Normal distribution with mean 0 and constant variance σ^2

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

- This is the same as

$$y_{ij} \sim N(\mu_i, \sigma^2)$$

STA 210

Analysis of Variance (ANOVA)

Main Idea: Decompose the total variation in the data into the variation between groups (model) and the variation within each group (residuals)

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

$$R^2 = \frac{\text{Variation between groups (model)}}{\text{Total variation}} = \frac{\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2}$$

STA 210

Total Variation

- Total variation = variation between and within groups

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

- Degrees of freedom

$$DFT = n - 1$$

- Estimate of the variance across all observations:

$$\frac{SST}{DFT} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2}{n - 1} = s_y^2$$

STA 210

Between Variation (Model)

- Variation in the group means

$$SSB = \sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2$$

- **Degrees of freedom**

$$DFB = K - 1$$

- **Mean Squares Between**

$$MSB = \frac{SSB}{DFB} = \frac{\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2}{K - 1}$$

- MSB is an estimate of the variance of the μ_i's

Within Variation (Residual)

- Variation within each group

$$SSW = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_k)^2$$

- **Degrees of freedom**

$$DFW = n - K$$

- **Mean Squares Within**

$$MSW = \frac{SSW}{DFW} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n - K}$$

- MSW is the estimate of σ^2, the variance within each group

STA 210

Using ANOVA to test difference in means

- **Question of interest** Is the mean value of the response y the same for all groups, or is there at least one group with a significantly different mean value?

- To answer this question, we will test the following hypotheses:

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = \cdots = \mu_K \\ &H_a : \text{At least one } \mu_i \text{ is not equal to the others} \end{aligned}$$

- **How to think about it:** If the sample means are "far apart", " there is evidence against H_0

- We will calculate a test statistic to quantify "far apart" in the context of the data

Analysis of Variance (ANOVA)

- **Main Idea:** Decompose the total variation in the data into the variation between groups and the variation within each group

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{K} n_i(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

- If the variation between groups is significantly greater than the variation within each group, then there is evidence against the null hypothesis.

ANOVA table for comparing means

	Sum of Squares	DF	Mean Square	F-Stat	p-value
Between (Model)	$\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2$	$K-1$	$SSB/(K-1)$	MSB/MSW	$P(F > \text{F-Stat})$
Within (Residual)	$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	$n-K$	$SSW/(n-K)$		
Total	$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$	$n-1$	$SST/(n-1)$		

Using ANOVA to test difference in means

Hypotheses:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_K \\ H_a &: \text{At least one } \mu_i \text{ is not equal to the others} \end{aligned}$$

Test statistic:

$$\frac{MSB}{MSW} = \frac{\sum\limits_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2/(K-1)}{\sum\limits_{i=1}^{K}\sum\limits_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(n-K)}$$

STA 210

Calculate p-value

Calculate the p-value using an F distribution with $K - 1$ and $n - K$ degrees of freedom

Capital Bike Share: ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
season	3	128202929	42734310	14.458	0
Residuals	96	283747246	2955700	NA	NA

- Go to http://bit.ly/sta210-sp20-anova and use the model to answer the questions
- Use **NetId@duke.edu** for your email address.
- You are welcome (and encouraged!) to discuss these questions with 1 - 2 people around you, but **each person** must submit a response.

04:00

Assumptions for ANOVA

- Normality: $y_{ij} \sim N(\mu_i, \sigma^2)$
- Constant variance: The population distribution for each group has a common variance, σ^2
- Independence: The observations are independent from one another
  - This applies to observation within and between groups
- We can typically check these assumptions in the exploratory data analysis

Robustness to Assumptions

- Normality: $y_{ij} \sim N(\mu_i, \sigma^2)$
  - ANOVA relatively robust to departures from Normality.
  - Concern when there are strongly skewed distributions with different sample sizes (especially if sample sizes are small, < 10 in each group)
- Independence: There is independence within and across groups
  - If this doesn't hold, should use methods that account for correlated errors

Robustness to Assumptions

- Constant variance: The population distribution for each group has a common variance, σ^2
    - Critical assumption, since the pooled (combined) variance is important for ANOVA
    - **General rule:** If the sample sizes within each group are approximately equal, the results of the F-test are valid if the largest variance is no more than 4 times the smallest variance (i.e. the largest standard deviation is no more than 2 times the smallest standard deviation)

STA 210

Capital Bike Share: Normality

```
ggplot(data = bikeshare, aes(x = count)) +
  geom_histogram() +
  facet_wrap(~season) +
  labs(title = "Daily bike rentals by season")
```

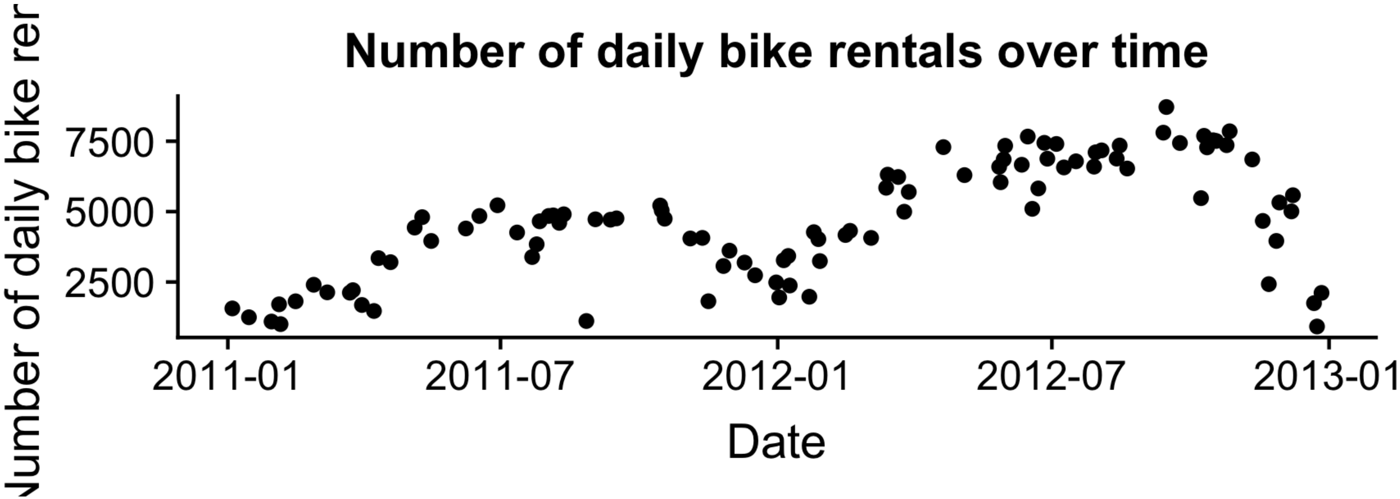
Capital Bike Share: Constant Variance

```
bikeshare %>%
  group_by(season) %>%
  summarise(sd = sd(count))
```

```
## # A tibble: 4 x 2
##   season    sd
##   <chr>  <dbl>
## 1 Fall   1848.
## 2 Spring 1889.
## 3 Summer 1662.
## 4 Winter 1465.
```

Capital Bike Share: Independence

- Recall that the data is 100 randomly selected days in 2011 and 2012.
- Let's look at the counts in date order to see if a pattern still exists



Though the days were randomly selected, it still appears the independence assumption is violated.

- Additional methods may be required to fully examine this data.

Why not just use the model output?

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	5180.200	343.843	15.066	0.000	4497.677	5862.723
seasonSpring	-256.591	496.726	-0.517	0.607	-1242.585	729.402
seasonSummer	558.911	477.178	1.171	0.244	-388.279	1506.101
seasonWinter	-2400.760	486.267	-4.937	0.000	-3365.993	-1435.527

- The model coefficients and associated hypothesis test / confidence interval are interpreted in relation to the baseline level
    - The coefficients, test statistics, confidence intervals, and p-values all change if the baseline category changes (more on this later!)
- An ANOVA test gives indication if any category has a significantly different mean regardless of the baseline
    - The sum of squares, mean squares, test statistic, and p-value stay the same even if the baseline changes

STA 210