

Analysis of

June 23, 2018

```
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##   filter, lag
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
## Loading required package: tidyr
## Loading required package: knitr
## Loading required package: ggplot2
## Loading required package: maps
## Loading required package: RColorBrewer
## Loading required package: summarytools
## Loading required package: magrittr
##
## Attaching package: 'magrittr'
## The following object is masked from 'package:tidyr':
##
##   extract
## Loading required package: stargazer
##
## Please cite as:
## Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
## R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

Import Breweries Data

```
#import breweries data
breweries_data <- read.csv("../data/Breweries.csv", header=TRUE)

#TODO: all columns to lowercase
#colnames(breweries_data) %<>% tolower # column names to lower case. %<>% is a compound operator that a

#summary of breweries raw data
brewery_summary_raw <- select(breweries_data, State, Brew_ID) %>% #select columns
  dplyr::group_by(State) %>% #group by
  dplyr::summarize_all(funs(count=n_distinct(.), min(.), max(.), mean(.), median(.), s
```

```
#print the summary in a way that it doesn't look like vomit
kable(brewery_summary_raw, digits = 2)
```

State	count	min	max	mean	median	sd
AK	7	103	558	366.14	454.0	167.33
AL	3	287	479	393.00	413.0	97.55
AR	2	140	260	200.00	200.0	84.85
AZ	11	31	550	306.36	233.0	191.48
CA	39	4	556	281.92	311.0	178.40
CO	47	7	552	320.13	387.0	168.49
CT	8	90	513	271.88	242.5	162.87
DC	1	228	228	228.00	228.0	NaN
DE	2	317	540	428.50	428.5	157.68
FL	15	68	528	356.07	379.0	135.54
GA	7	50	476	325.43	401.0	158.69
HI	4	204	440	306.25	290.5	120.36
IA	5	209	483	399.80	469.0	116.69
ID	5	170	314	264.20	308.0	65.98
IL	18	41	553	152.44	70.5	143.05
IN	22	17	507	128.82	27.5	147.27
KS	3	46	501	277.00	284.0	227.58
KY	4	2	389	138.75	82.0	171.74
LA	5	153	554	337.20	270.0	193.96
MA	23	3	512	277.52	294.0	131.19
MD	7	69	522	253.86	256.0	186.93
ME	9	43	503	303.78	318.0	154.54
MI	32	8	542	169.06	123.5	159.55
MN	12	1	475	186.33	141.0	155.85
MO	9	32	443	224.11	189.0	158.24
MS	2	134	246	190.00	190.0	79.20
MT	9	220	544	444.89	500.0	111.24
NC	19	70	541	333.84	360.0	162.77
ND	1	336	336	336.00	336.0	NaN
NE	5	190	509	340.20	338.0	118.32
NH	3	48	548	235.33	110.0	272.55
NJ	3	218	282	241.00	223.0	35.59
NM	4	266	444	359.00	363.0	76.82
NV	2	234	531	382.50	382.5	210.01
NY	16	47	557	360.56	336.5	155.06
OH	15	92	435	211.07	184.0	118.76
OK	6	183	506	326.33	349.0	122.13
OR	29	81	495	285.45	207.0	132.74
PA	25	44	546	291.40	323.0	143.94
RI	5	87	380	198.60	144.0	126.79
SC	4	6	519	289.25	316.0	219.07
SD	1	213	213	213.00	213.0	NaN
TN	3	240	520	357.67	313.0	145.25
TX	28	30	471	210.32	186.5	124.15
UT	4	160	400	307.00	334.0	105.89
VA	16	51	456	294.62	325.5	126.02
VT	10	42	502	276.50	274.5	117.27


```

# capture potential duplicates
breweries_dups <- filter(breweries_summary, Brew_ID_count > 1) # if Brew_ID_count > 1 then there is a p

# rejoin potential dups to original dataset
breweries_dups <- select(breweries_dups %>% inner_join(breweries_data, by="Name"), -ends_with("_count"))

# Fix Errors #

# Fix Brew_ID=378, change City(Menominee -> Menominie)
breweries_dups <- breweries_dups %>%
  mutate(City=replace(City, Brew_ID==378, "Menominie")) %>%
  as.data.frame()

# Fix Brew_ID=96, change State(MA -> MI)
breweries_dups <- breweries_dups %>%
  mutate(State=replace(State, Brew_ID==96, "MI")) %>%
  as.data.frame()

#capture known duplicates
breweries_dups <- breweries_dups %>%
  group_by(Name, City, State) %>%
  filter(n()>1)

#create surrogate key for duplicates
breweries_sk <- breweries_dups %>%
  group_by(Name, City, State) %>%
  summarize_all(funs(
    Brew_SK = (sum(Brew_ID)*sum(Brew_ID)),
    count = n()
  )) %>% #end summarize_all
  ungroup() %>%
  right_join(breweries_dups, by = c("Name", "City", "State")) %>% # rejoin to dupes b
  select(Brew_ID, Brew_SK)

breweries_data$Brew_ID[(breweries_data$Brew_ID %in% breweries_sk$Brew_ID)] <- breweries_sk$Brew_SK # up

breweries_clean <- distinct(breweries_data, Brew_ID, .keep_all = TRUE) %>% rename(Brewery_Name = Name)

#Check for Outliers
#Impute missing values

summary(breweries_clean)

```

```

##      Brew_ID      Brewery_Name      City
## Min.      :    1.0  Blackrocks Brewery :    2  Length:555
## 1st Qu.:   143.5  Blue Mountain Brewery :    2  Class :character
## Median :   282.0  Oskar Blues Brewery   :    2  Mode  :character
## Mean      :  1627.3  Otter Creek Brewing   :    2
## 3rd Qu.:   421.5  Sly Fox Brewing Company :    2

```

```
## Max.      :697225.0    10 Barrel Brewing Company: 1
##              (Other)              :544
##      State
## Length:555
## Class :character
## Mode  :character
##
##
##
##
```

```
# See stats.rmd
```

Clean Beer Data

```
beer_data <- read.csv("../data/Beers.csv", header=TRUE)
```

```
head(beer_data)
```

```
##           Name Beer_ID  ABV IBU Brewery_id
## 1      Pub Beer   1436 0.050  NA       409
## 2    Devil's Cup   2265 0.066  NA       178
## 3 Rise of the Phoenix 2264 0.071  NA       178
## 4      Sinister   2263 0.090  NA       178
## 5    Sex and Candy  2262 0.075  NA       178
## 6    Black Exodus  2261 0.077  NA       178
##           Style Ounces
## 1    American Pale Lager      12
## 2    American Pale Ale (APA)    12
## 3          American IPA        12
## 4 American Double / Imperial IPA 12
## 5          American IPA        12
## 6      Oatmeal Stout          12
```

```
beer_data$Brewery_id[(beer_data$Brewery_id %in% breweries_sk$Brew_ID)] <- breweries_sk$Brew_SK # update
```

```
## Warning in beer_data$Brewery_id[(beer_data$Brewery_id %in% breweries_sk
## $Brew_ID)] <- breweries_sk$Brew_SK: number of items to replace is not a
## multiple of replacement length
```

```
beer_clean <- distinct(beer_data) %>% rename(Brew_ID = Brewery_id, Beer_Name = Name) #
```

```
# kable(as.data.frame(summarytools::descr(beer_clean)), digits = 2)
```

Question 1

```
state_ll <- read.csv("../data/state_coords.csv") %>% mutate(State = toupper(State)) %>% rename(state = State)
states <- map_data("state") %>%
  mutate(region = toupper(region)) %>%
  rename(state=region) %>%
  select(long, lat, state, group)
```

```

states <- states %>%
  left_join(
    states %>%
      group_by(state) %>%
      summarise_all(funs(n=n())) %>%
      select(state, group_n) %>%
      distinct(state, .keep_all = TRUE)
  )

```

```
## Joining, by = "state"
```

```

breweries_by_state <- select(breweries_clean, Brew_ID, State) %>%
  group_by(State) %>%
  summarise_all(funs(Brewery_count = n()))

```

```

# state_ll %>%
#   inner_join(states)

```

```
kable(as.data.frame(summarytools::descr(breweries_by_state, transpose = TRUE)), digits = 2)
```

	Mean	Std.Dev	Min	Median	Max	MAD	IQR	CV	Skewness	SE.Skewness	Kurtosis	N
Brewery_count	10.88	10.59	1	7	47	5.93	12.5	1.03	1.43	0.33	1.57	

```
freq(breweries_clean$State, order = "freq")
```

```
## Frequencies
```

```
## State
```

```
## Data frame: breweries_clean
```

```
## Type: Character
```

```
##
```

```

##          Freq  % Valid  % Valid Cum.  % Total  % Total Cum.
## -----
##          CO    47     8.47      8.47     8.47     8.47
##          CA    39     7.03     15.50     7.03     15.50
##          MI    32     5.77     21.26     5.77     21.26
##          OR    29     5.23     26.49     5.23     26.49
##          TX    28     5.05     31.53     5.05     31.53
##          PA    25     4.50     36.04     4.50     36.04
##          MA    23     4.14     40.18     4.14     40.18
##          WA    23     4.14     44.32     4.14     44.32
##          IN    22     3.96     48.29     3.96     48.29
##          NC    19     3.42     51.71     3.42     51.71
##          WI    19     3.42     55.14     3.42     55.14
##          IL    18     3.24     58.38     3.24     58.38
##          NY    16     2.88     61.26     2.88     61.26
##          VA    16     2.88     64.14     2.88     64.14
##          FL    15     2.70     66.85     2.70     66.85
##          OH    15     2.70     69.55     2.70     69.55
##          AZ    11     1.98     71.53     1.98     71.53
##          MN    10     1.80     73.33     1.80     73.33
##          VT    10     1.80     75.14     1.80     75.14

```

##	ME	9	1.62	76.76	1.62	76.76
##	MO	9	1.62	78.38	1.62	78.38
##	MT	9	1.62	80.00	1.62	80.00
##	CT	8	1.44	81.44	1.44	81.44
##	AK	7	1.26	82.70	1.26	82.70
##	GA	7	1.26	83.96	1.26	83.96
##	MD	7	1.26	85.23	1.26	85.23
##	OK	6	1.08	86.31	1.08	86.31
##	IA	5	0.90	87.21	0.90	87.21
##	ID	5	0.90	88.11	0.90	88.11
##	LA	5	0.90	89.01	0.90	89.01
##	NE	5	0.90	89.91	0.90	89.91
##	RI	5	0.90	90.81	0.90	90.81
##	HI	4	0.72	91.53	0.72	91.53
##	KY	4	0.72	92.25	0.72	92.25
##	NM	4	0.72	92.97	0.72	92.97
##	SC	4	0.72	93.69	0.72	93.69
##	UT	4	0.72	94.41	0.72	94.41
##	WY	4	0.72	95.14	0.72	95.14
##	AL	3	0.54	95.68	0.54	95.68
##	KS	3	0.54	96.22	0.54	96.22
##	NH	3	0.54	96.76	0.54	96.76
##	NJ	3	0.54	97.30	0.54	97.30
##	TN	3	0.54	97.84	0.54	97.84
##	AR	2	0.36	98.20	0.36	98.20
##	DE	2	0.36	98.56	0.36	98.56
##	MS	2	0.36	98.92	0.36	98.92
##	NV	2	0.36	99.28	0.36	99.28
##	DC	1	0.18	99.46	0.18	99.46
##	ND	1	0.18	99.64	0.18	99.64
##	SD	1	0.18	99.82	0.18	99.82
##	WV	1	0.18	100.00	0.18	100.00
##	<NA>	0			0.00	100.00
##	Total	555	100.00	100.00	100.00	100.00

```

#map of breweries by state
ggplot(data = breweries_by_state %>%
  inner_join(state_ll, by=c("State" = "Abbr")) %>%
  inner_join(states)) +
  geom_polygon(aes(x = long, y = lat, group=group, fill=Brewery_count), color = "black") +
  #geom_text(aes(x = long, y = lat, label = as.character(Brewery_count), color = "black")) +
  coord_fixed(1.3) +
  guides(alpha=FALSE)

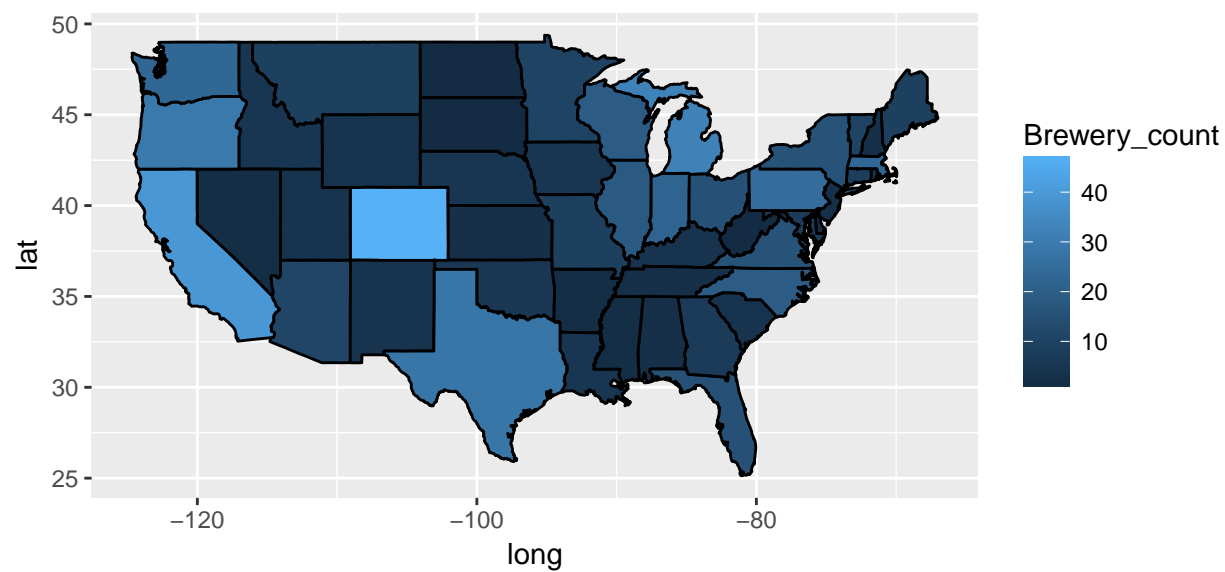
```

```

## Warning: Column `State`/`Abbr` joining character vector and factor,
## coercing into character vector

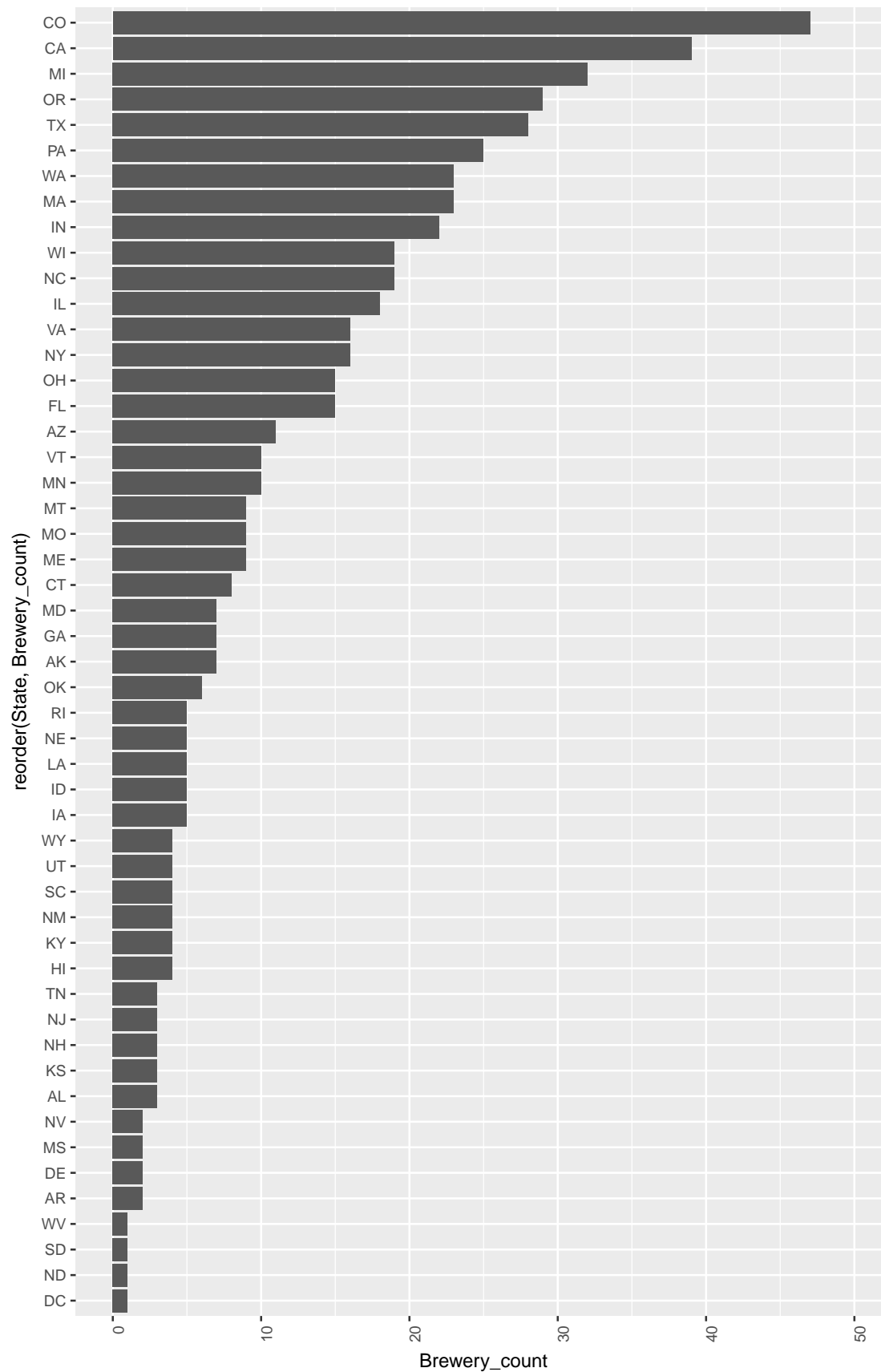
## Joining, by = "state"

```



```
# scale_fill_gradientn(colours = "black",
#                       breaks = c(2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50))

#barplot brewery data by state
ggplot(breweries_by_state, aes(x=reorder(State, Brewery_count), y= Brewery_count)) + #TODO: Make Pretty
  geom_bar(stat="identity") +
  ylim(0, 50) +
  theme(text = element_text(size=10),
        axis.text.x = element_text(angle=90, hjust=1)) +
  scale_fill_hue(c=45, l=40, direction = -1) +
  coord_flip()
```

Question 2

```
# merge beer and breweries
merged_data <- breweries_clean %>%
  full_join(beer_clean, by="Brew_ID")

#TODO: Plot -> brews by brewery
```

Question 3

```
# Number of nulls in each column
merged_data %>%
  select_if(function(x) any(is.na(x))) %>%
  summarise_all(funs(sum(is.na(.))))
```

```
##   ABV  IBU
## 1   62 1005
```

```
#TODO: add plot?
```

Question 4

```
#fig.width=11

# compute median ABV and IBU by state

merged_by_state <- select(merged_data, State, ABV, IBU) %>%
  group_by(State) %>%
  summarise_all(median)#funs(median(!is.na(.)))) #TODO: Double check this is calculating median

merged_by_state$State <- as.factor(merged_by_state$State)

# vars <- rbind(merged_by_state %>% mutate(var="ABV") %>%rename(value=ABV) %>% select(State, var, value),
#               merged_by_state %>% mutate(var="IBU") %>%rename(value=IBU) %>% select(State, var, value))
#
# vars$value

summary(merged_by_state)
```

```
##      State      ABV      IBU
## AK      : 1  Min.   :0.04000  Min.   :32.00
## AL      : 1  1st Qu.:0.05400  1st Qu.:33.88
## AR      : 1  Median :0.05550  Median :39.75
## AZ      : 1  Mean    :0.05514  Mean    :42.25
## CA      : 1  3rd Qu.:0.05800  3rd Qu.:48.12
## CO      : 1  Max.    :0.06250  Max.    :57.50
## (Other):45  NA's    :18      NA's    :47
```

```
kable(as.data.frame(summarytools::descr(beer_clean)), digits = 2)
```

	Beer_ID	ABV	IBU	Brew_ID	Ounces
Mean	1431.11	0.06	42.71	1772.99	13.59
Std.Dev	752.46	0.01	25.95	31761.76	2.35
Min	1.00	0.00	4.00	1.00	8.40
Median	1453.50	0.06	35.00	207.00	12.00
Max	2692.00	0.13	138.00	697225.00	32.00
MAD	934.78	0.01	25.20	194.22	0.00
IQR	1267.50	0.02	43.00	273.50	4.00
CV	1.90	4.41	1.65	0.06	5.78
Skewness	-0.12	0.96	0.79	21.78	2.04
SE.Skewness	0.05	0.05	0.07	0.05	0.05
Kurtosis	-1.09	1.14	-0.14	473.88	9.01
N.Valid	2410.00	2348.00	1405.00	2410.00	2410.00
Pct.Valid	100.00	97.43	58.30	100.00	100.00

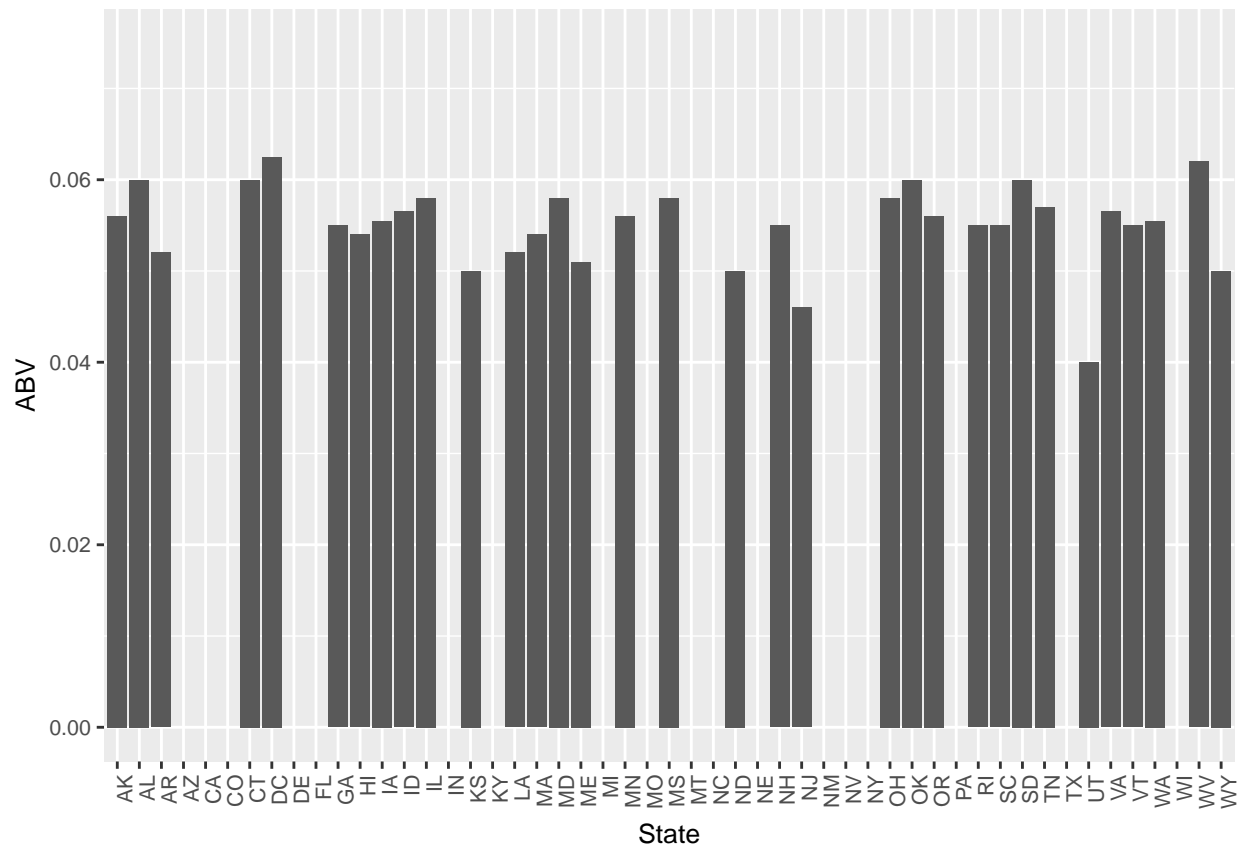
```
merged_by_state %>% na.omit(IBU)
```

```
## # A tibble: 4 x 3
##   State    ABV    IBU
##   <fctr> <dbl> <dbl>
## 1 MS     0.0580  45.0
## 2 ND     0.0500  32.0
## 3 NJ     0.0460  34.5
## 4 WV     0.0620  57.5
```

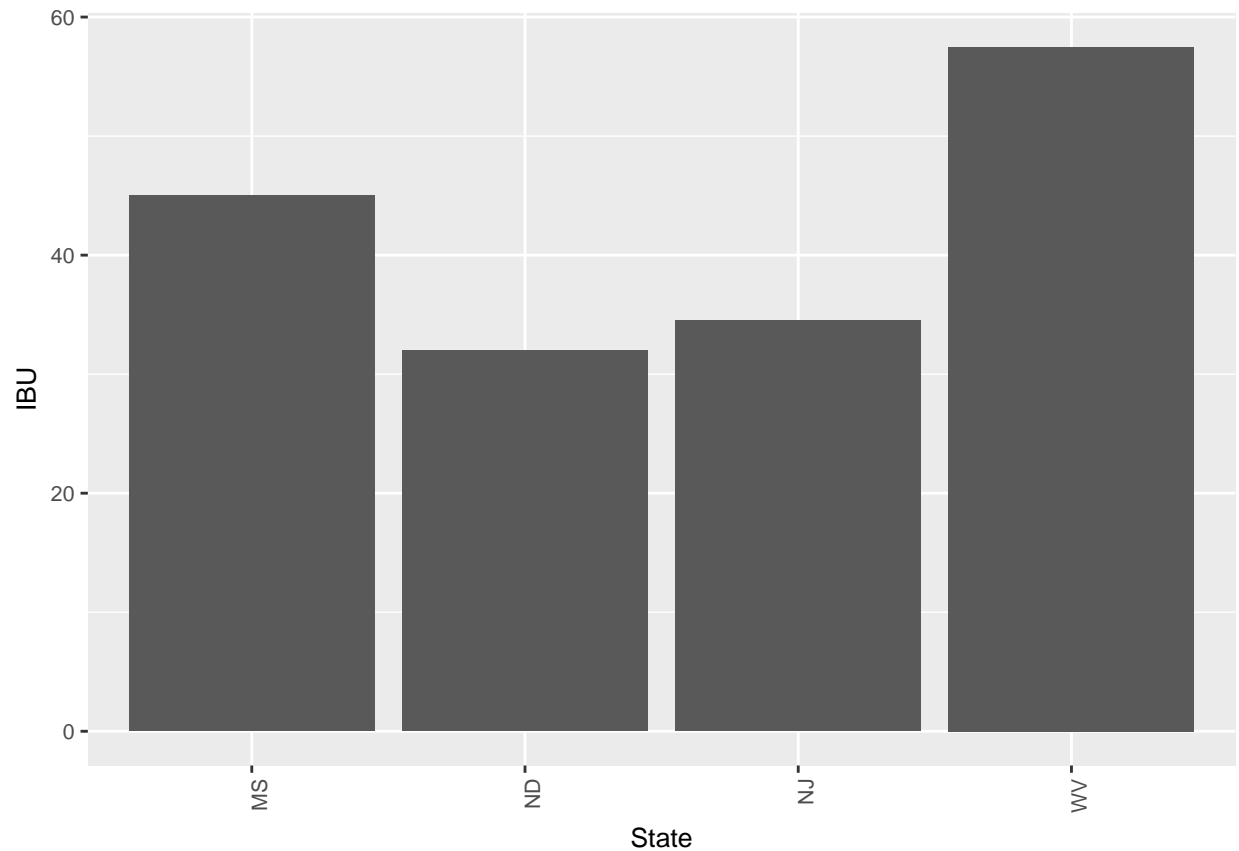
```
#MEDIAN
```

```
ggplot(merged_by_state, aes(x=State, y=ABV)) +
  geom_bar(stat = "identity", position = "dodge") +
  ylim(0, .075) +
  theme(text = element_text(size=10),
        axis.text.x = element_text(angle=90, hjust=1))
```

```
## Warning: Removed 18 rows containing missing values (geom_bar).
```



```
ggplot((merged_by_state %>% na.omit()), aes(x=State, y=IBU)) + #TODO: something is fishy with IBU
  geom_bar(stat = "identity", position = "dodge") +
  #ylim(0, .075) +
  theme(text = element_text(size=10),
        axis.text.x = element_text(angle=90, hjust=1))
```

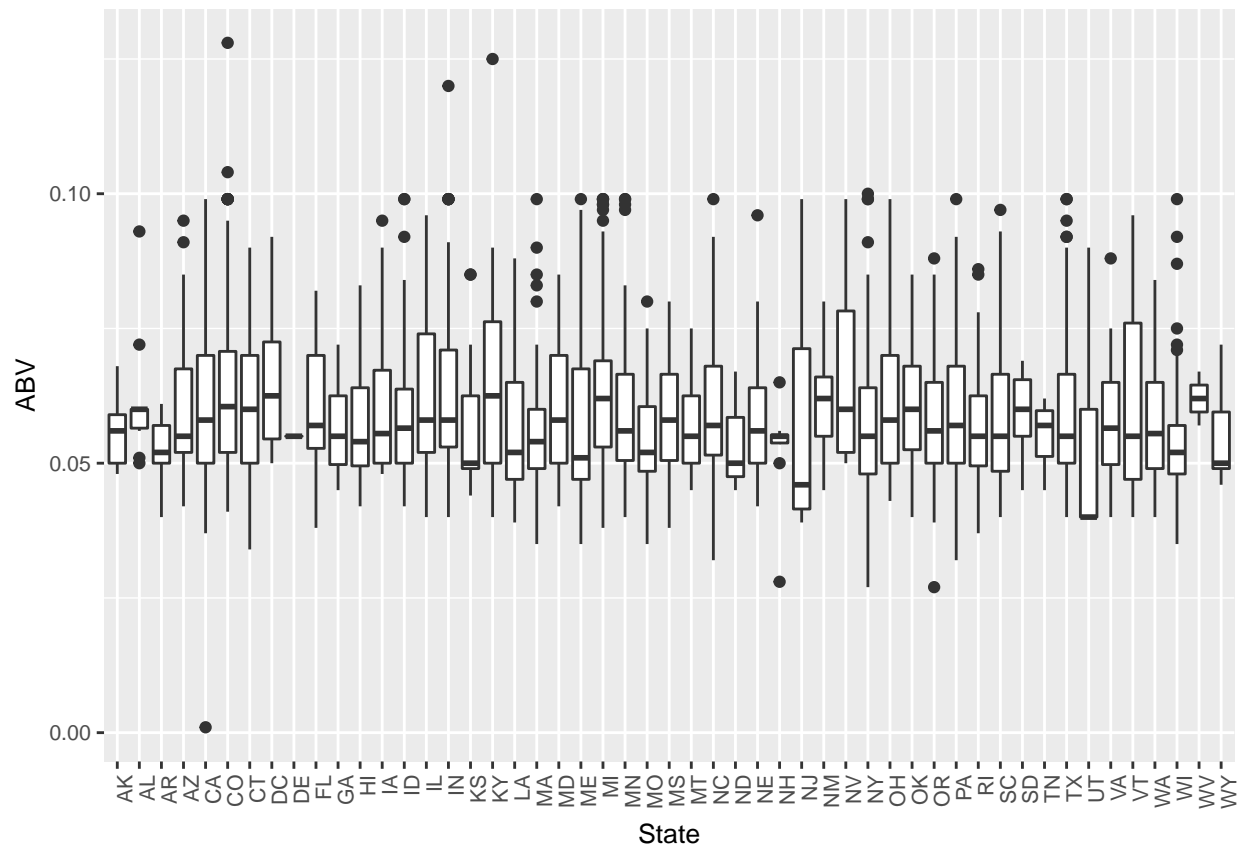


Question 5

```
# max_abv <- max(merged_data$ABV, na.rm = TRUE)

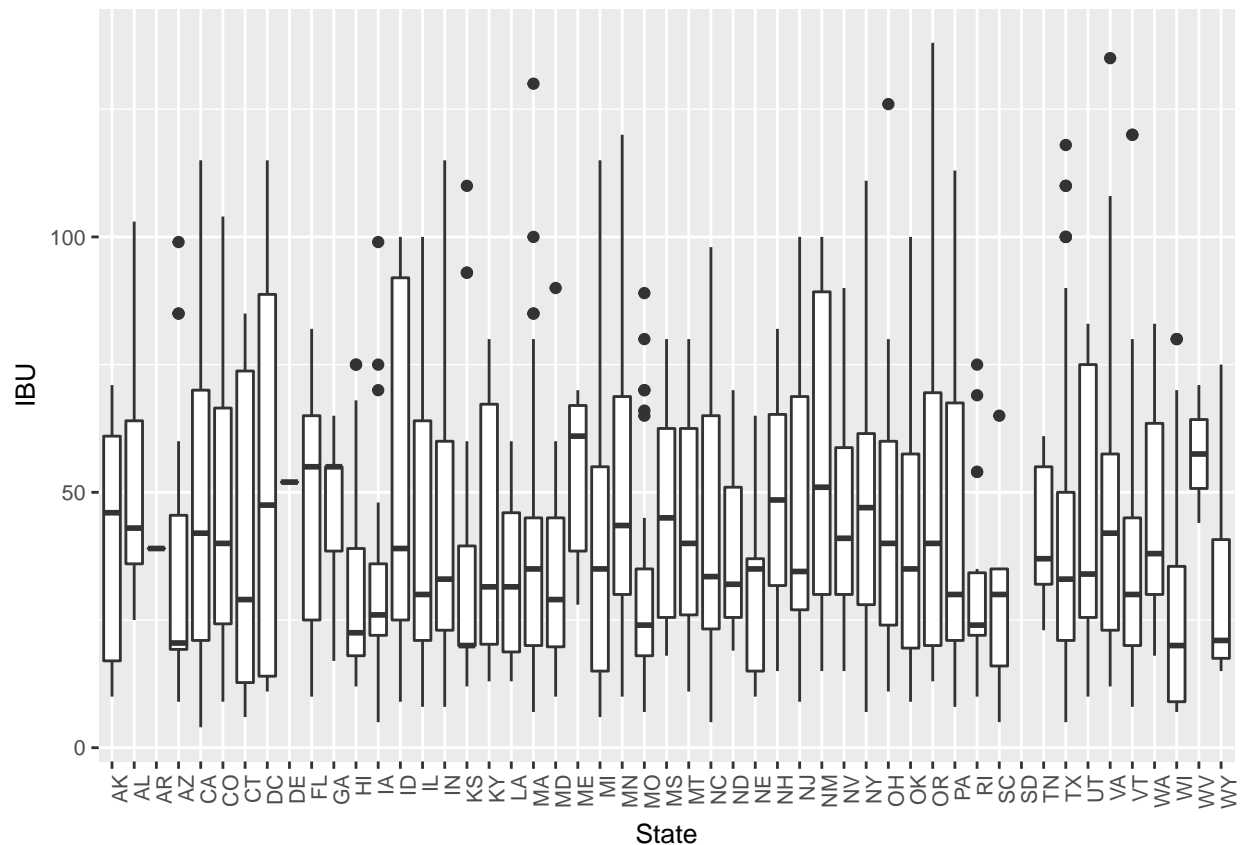
ggplot(merged_data, aes(x=State , y=ABV)) + #TODO: Make Pretty
  geom_boxplot() +
  #ylim(0, .075) +
  theme(text = element_text(size=10),
        axis.text.x = element_text(angle=90, hjust=1))
```

```
## Warning: Removed 62 rows containing non-finite values (stat_boxplot).
```



```
ggplot(merged_data, aes(x=State , y=IBU)) + #TODO: Make Pretty
  geom_boxplot() +
  #ylim(0, .075) +
  theme(text = element_text(size=10),
        axis.text.x = element_text(angle=90, hjust=1))
```

```
## Warning: Removed 1005 rows containing non-finite values (stat_boxplot).
```



```
max_abv <- (select(merged_data, State, ABV) %>%
  group_by(State) %>%
  #filter(ABV == max(ABV)) %>%
  arrange(desc(ABV)) %>% #sort by ABV
  filter(row_number() == 1))[1,] #get first row
```

```
max_abv
```

```
## # A tibble: 1 x 2
## # Groups:   State [1]
##   State ABV
##   <chr> <dbl>
## 1 CO    0.128
```

```
max_ibu <- (select(merged_data, State, IBU) %>%
  group_by(State) %>%
  #filter(ABV == max(ABV)) %>%
  arrange(desc(IBU)) %>% #sort by ABV
  filter(row_number() == 1))[1,] #get first row
```

```
max_ibu
```

```
## # A tibble: 1 x 2
## # Groups:   State [1]
##   State IBU
##   <chr> <int>
## 1 OR    138
```

Question 6

```
#summaryize ABV

# tidy_summary <- tidy(summary(merged_data$ABV)) #For some reason this line wont knit

abv_stats <- as.data.frame(t(summary(merged_data$ABV))) %>% #summarize and transpose
  rename("ABV"=Freq, Statistic=Var2) %>%
  select(Statistic, ABV)

abv_stats$ABV <- round(abv_stats$ABV, digits = 3)

abv_stats #TODO: Add IQR, stdev      #TODO: Compare to quinton's summary

##   Statistic   ABV
## 1      Min. 0.001
## 2    1st Qu. 0.050
## 3     Median 0.056
## 4       Mean 0.060
## 5    3rd Qu. 0.067
## 6        Max. 0.128
## 7        NA's 62.000
```

Question 7

```
# fig.height=48
#plot relationship of ABV and IBU

#retrieve linear model equation -- source(https://stackoverflow.com/questions/7549694/adding-regression)
lm_eqn = function(m) {

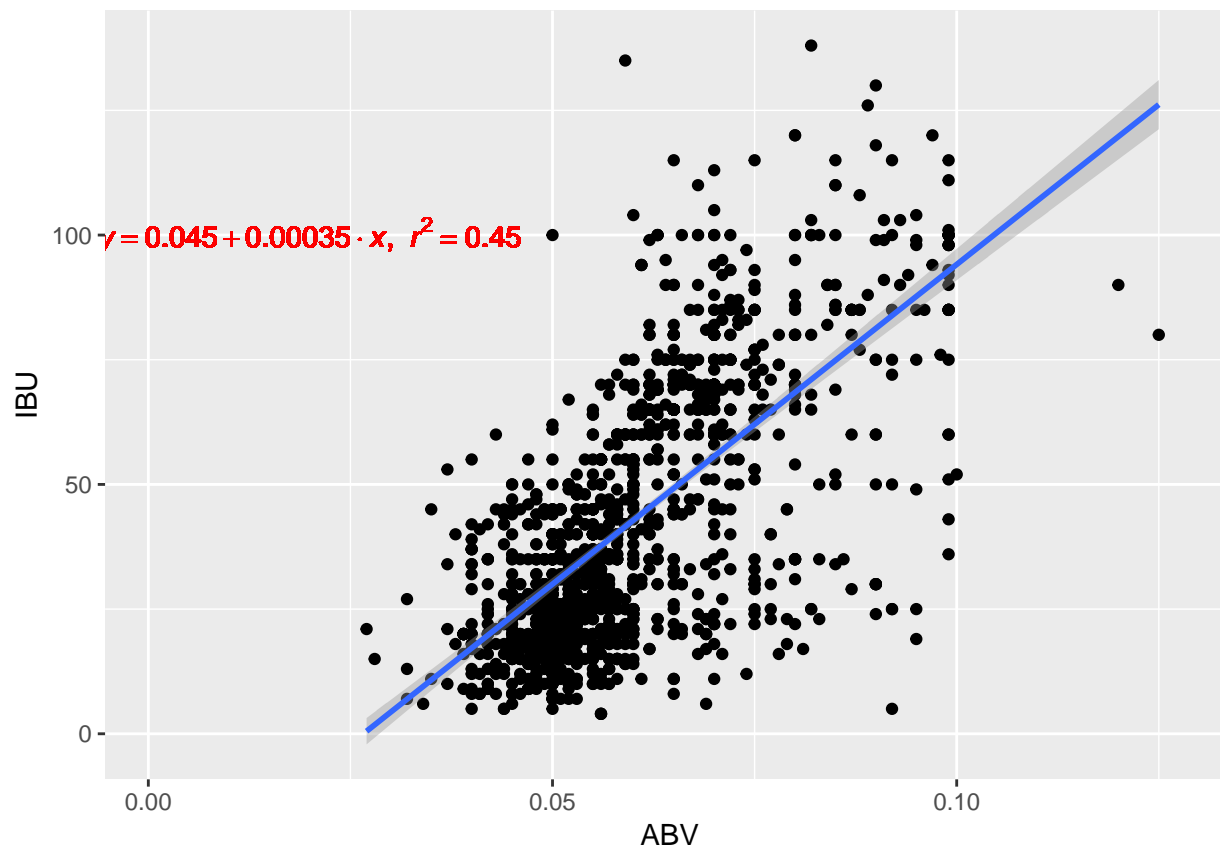
  l <- list(a = format(coef(m)[1], digits = 2),
           b = format(abs(coef(m)[2]), digits = 2),
           r2 = format(summary(m)$r.squared, digits = 3));

  if (coef(m)[2] >= 0) {
    eq <- substitute(italic(y) == a + b %.% italic(x)*", "~italic(r)^2~"=~r2,l)
  } else {
    eq <- substitute(italic(y) == a - b %.% italic(x)*", "~italic(r)^2~"=~r2,l)
  }

  as.character(as.expression(eq));
}

ggplot(beer_clean, aes(x=ABV, y=IBU)) +
  geom_point() +
  geom_smooth(method = "lm") +
  geom_text(aes(x = .02, y = 100, label = lm_eqn(lm(ABV ~ IBU ,beer_clean))), parse = TRUE, color = "red")

## Warning: Removed 1005 rows containing non-finite values (stat_smooth).
## Warning: Removed 1005 rows containing missing values (geom_point).
```

Yes, there is a positive relationship between ABV and IBU. #TODO:Add explanation

Appendix

Session Info

```
sessionInfo()
```

```
## R version 3.4.3 (2017-11-30)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 16299)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## other attached packages:
## [1] bindrcpp_0.2      stargazer_5.2      magrittr_1.5
```

```

## [4] summarytools_0.8.0   RColorBrewer_1.1-2   maps_3.2.0
## [7] ggplot2_2.2.1        knitr_1.18           tidyr_0.7.2
## [10] dplyr_0.7.4          RevoUtilsMath_10.0.1 RevoUtils_10.0.7
## [13] RevoMods_11.0.0      MicrosoftML_9.3.0    mrsdeploy_1.1.3
## [16] RevoScaleR_9.3.0     lattice_0.20-35      rpart_4.1-11
##
## loaded via a namespace (and not attached):
## [1] purrr_0.2.4          pander_0.6.1          colorspace_1.3-2
## [4] htmltools_0.3.6      yaml_2.1.16           CompatibilityAPI_1.1.0
## [7] utf8_1.1.2           rlang_0.1.6           pillar_1.0.1
## [10] glue_1.2.0           pryr_0.1.3            matrixStats_0.52.2
## [13] foreach_1.4.5        bindr_0.1             plyr_1.8.4
## [16] stringr_1.2.0         munsell_0.4.3         gtable_0.2.0
## [19] codetools_0.2-15     evaluate_0.10.1       labeling_0.3
## [22] curl_3.1             highr_0.6             Rcpp_0.12.14
## [25] scales_0.5.0         backports_1.1.2       jsonlite_1.5
## [28] rapporttools_1.0     digest_0.6.13         stringi_1.1.6
## [31] grid_3.4.3           rprojroot_1.3-1       cli_1.0.0
## [34] tools_3.4.3          bitops_1.0-6          lazyeval_0.2.1
## [37] RCurl_1.95-4.9       tibble_1.4.1          crayon_1.3.4
## [40] pkgconfig_2.0.1      assertthat_0.2.0      rmarkdown_1.8
## [43] iterators_1.0.9      R6_2.2.2              compiler_3.4.3

```