R_Final

Kalenderoglu, Ugurcan

6 Ocak 2019

Part 1

1.R or Python?

Debate on Python vs R is one of thing that has no true answer i believe. Both sofwares have wide range of applications, both are open source and there are tons of documents, discussions and shared bug-fixing soluntions available on the internet for both. Although it's must to master at least one of two, there are more crucial things will benefit you more than concluding trade-off between R and python if you want to be Data Scientist.

I've read an article mentioned and totally agree. I think there is an illusion of competence; once you've mastered on one of them, you begin to think that it is superior over other. I would opt to accept that each have more advantageous in different part of the data science journey. R is really practical and handy when exploring, summarizing, visualizing the data while Python has wider usage in machine learning pipelines and more computing-heavy topics which require coding.So it's more safe to compare them case by case rather than stating bold idea such that R is better than Python at all.There is an also effect of developers' background who contribute to the development of libraries for them. For instance, R is widely used by statisticans compared to Python, since most of the R's developers are from statistics background.

To sum up, it's best to learn both softwares and use proper one depending on the task that you try to complete.

2.EDA Workflow

Exploratary data analysis should start with understanding of dimensions of the data which are number of rows, number of variables and types of each variable. Then it would be good to get summary statistics like mean, median, quartiles etc for numeric variables. Thirdly, visualization is generally best way to explore your data. If you have categorical variable, maybe histogram would give an meaningful idea as a starting point. So, choosing the relevant graph type and embedding right data into it would help much. As a last step, dealing with missing values and way of handling outliers are crucial since they can change what the data says to audience.

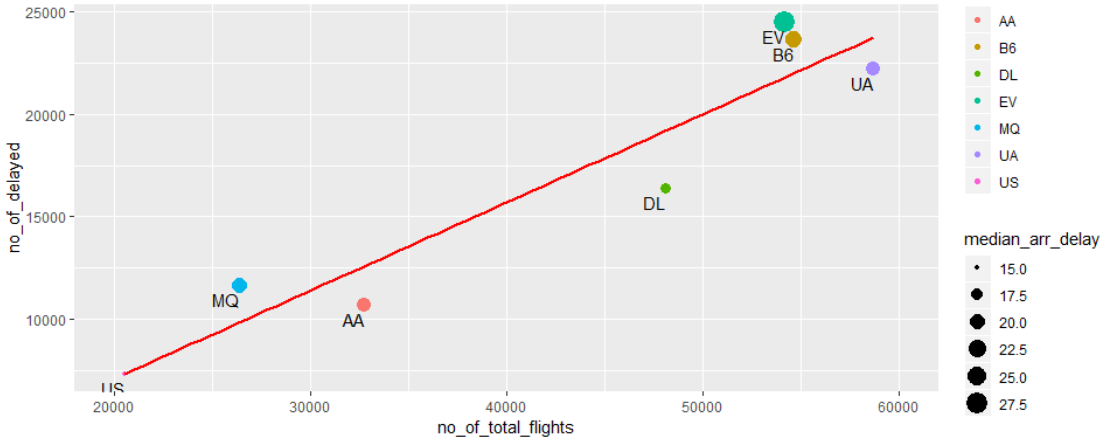
If i were expected to deliver insight on choosing optimal mix for investing donations, i would definitely start with definition of "positive impact on the society". I may transform it into measurable thing like how many people's life has changed as a result of activities of this projects by benchmarking similiar projects. Then i would try to gather the data on result of activities of these projects: how much money spent and how many people are

affected. I would definitely use the title starting with “pain points…”. Although the data is new oil, it should not be treated as “answer for all questions that humankind has”; we should still rely on human judgement at the decision stage with the help of data.

3.Flights Data

I would be interested to see how frequently carrier arrived with delay compared to total number of flights and median delay time will also be good indicator to comment on. Scatter plot would be best fit for that kind of visualization need. Putting no_of_total_flights on x axis and no_of_delayed_flights on y axis and also mapping carrier to colour, median_arr_delay to size of dots will be required steps to get this graph with ggplot2 package.Finally adding smooth line will make it easier to comment on best performer, worst performer in terms of timely arriving perspective.

```
#Get number of total flights per carrier
total_f <- flights %>%
  group_by(carrier) %>%
  summarize(no_of_total_flights = n())
#Get number of flights which arr_delay is bigger than 0 and median arr_delay
per carrier
total_d <- flights %>%
  filter(arr_delay > 0) %>%
  group_by(carrier) %>%
  summarize(no_of_delayed = n(),
            median_arr_delay = median(arr_delay))
#Join 2 tables and make it ready data for plotting
total_vf <- total_f %>%
  left_join(total_d, by = "carrier")
#Plot the Graph (In order to have well looking graph, we can eliminate
carries flew less than 20.000 times)
total_vf %>%
  filter(no_of_total_flights > 20000) %>%
  ggplot(aes(x=no_of_total_flights, y=no_of_delayed)) +
  scale_x_continuous(limits=c(20000,60000)) +
  geom_point(aes(color=carrier, size=median_arr_delay)) +
  geom_text(aes(label=carrier),hjust=1, vjust=1.5) +
  stat_smooth(method = "lm", se=F, col = "red")
```



It's more safe decision to buy ticket from UA and DL compared to EV and B6 when it's possible. UA and DL delayed below the expectation according to the smoothed average of benchmarks and also median delay time is lower.

Part 2: Extending the Group Project

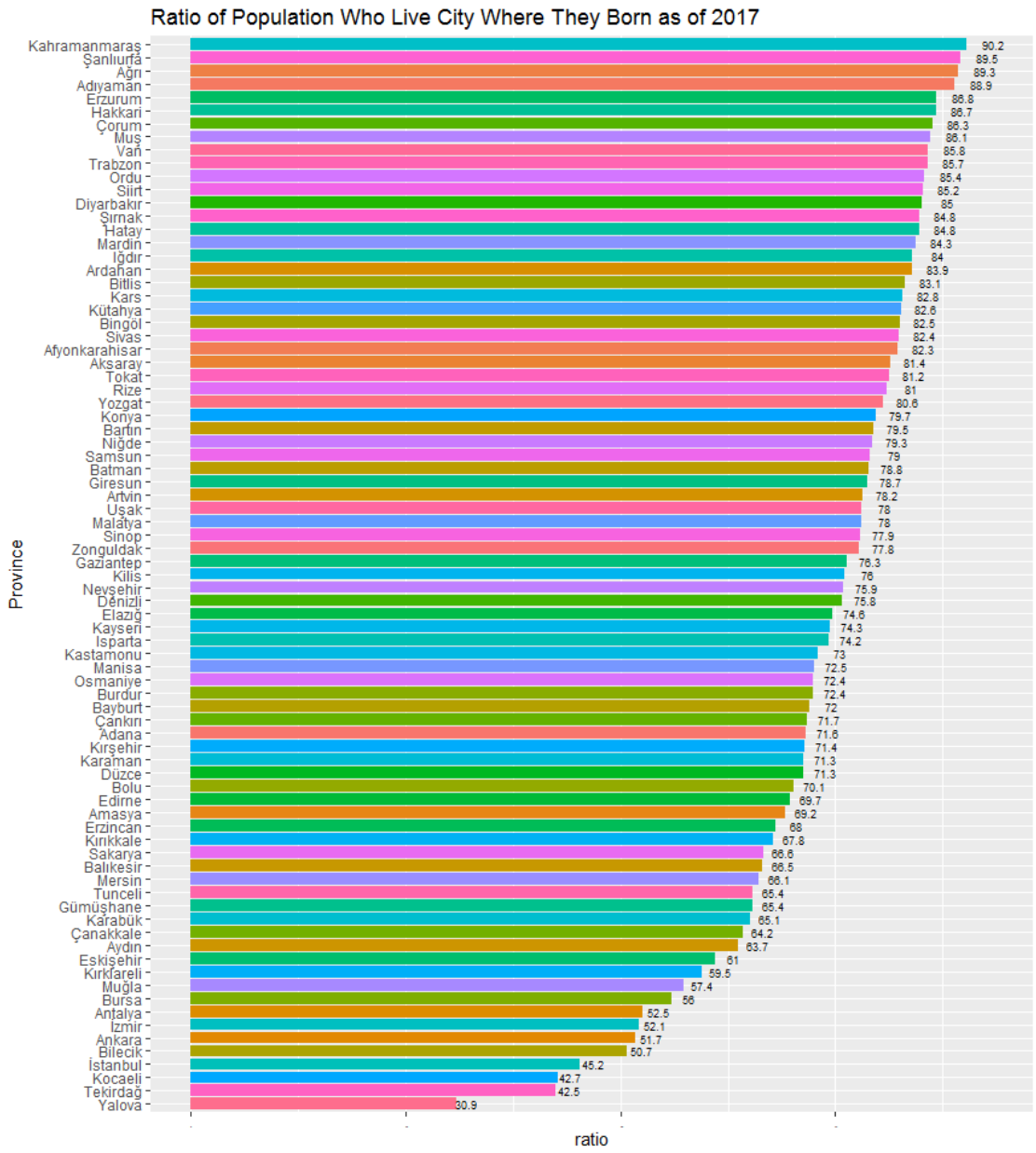
On top of previous analysis, it would be good perspective to create ratio matrice which shows that what percentage of the current population born in which city of Turkey. So we can reach conclusions such: -Top cities where majority of citizens have hometown different than city they live -Where do people who born abroad consist more percentage of population in the province and also less percentage of the total population.

```
#Read the Data
pop_data<-
read_excel("C:\\Users\\kalenderoglu\\Desktop\\BDA_assignments\\Group_Project\
\part_2\\origin_home_town_2016.xls")
#Melt & Format
pop_data <- melt(pop_data)

## Using X__1, Nüfusa Kayıtlı Olunan İl as id variables

colnames(pop_data) <- c("Year", "Province", "Birth_Place", "People")
pop_data <- pop_data %>%
  select(Province, Birth_Place, People)
pop_data$Birth_Place <- as.character(pop_data$Birth_Place)
pop_data$Province <- enc2native(pop_data$Province)
pop_data$Birth_Place <- enc2native(pop_data$Birth_Place)
#Arrange the Data & Mutate New Variable
ht_dist <- pop_data %>%
  group_by(Province) %>%
  mutate(ratio = round(People / sum(People) * 100,1)) %>%
  filter(Province == Birth_Place) %>%
  arrange(desc(ratio))
#Plot the Graph
ht_dist %>%
  ggplot(aes(x=reorder(Province, ratio), y=ratio, fill=Province, label =
```

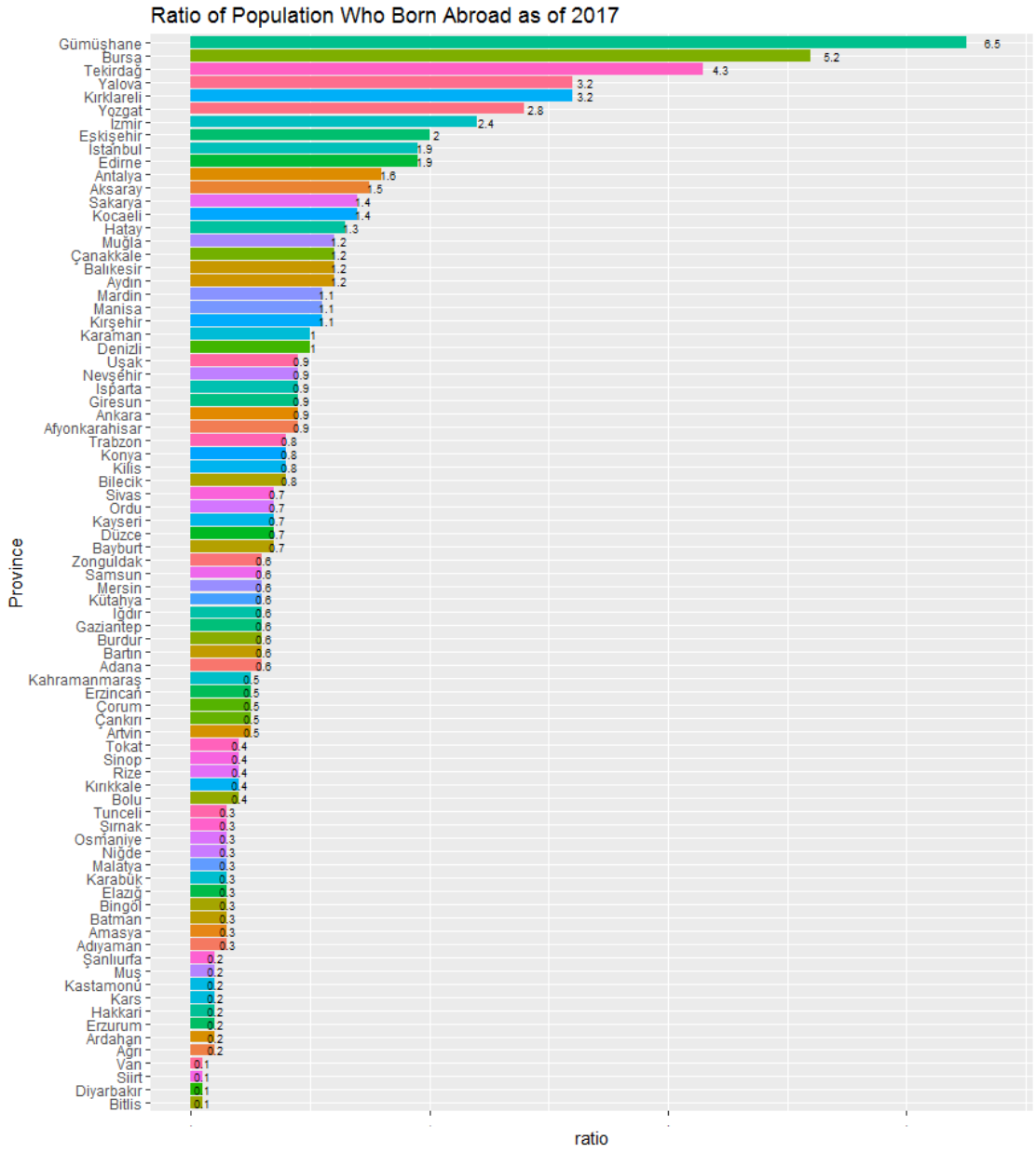
```
ratio)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(x = "Province", y = "ratio") +
  ggtitle("Ratio of Population Who Live City Where They Born as of 2017") +
  geom_text(size = 2.5, position = position_stack(vjust = 1.035)) +
  theme(legend.position = "none", axis.text.x = element_text(angle = 0.0,
vjust = 0.0, hjust = 0.0, size = 1))
```



Kahramanmaras is the top place where it's almost impossible to find anyone whose hometown is not K.Maras. It's surprising that most of the population live in Yalova and Bilecik born in different city than their current province while it's expected to see Istanbul,

Kocaeli and Tekirdag at the bottom of the graph. I'm also interested to explore the distribution of population born in abroad by city of province

```
#First, we should prepare the relevant data frame to fit into the model we
want
pop_abroad <- pop_data %>%
  group_by(Province) %>%
  mutate(ratio = round(People / sum(People) * 100,1)) %>%
  filter(Birth_Place == "Yurtdışı") %>%
  arrange(desc(ratio))
#Plot the Graph
pop_abroad %>%
  ggplot(aes(x=reorder(Province, ratio), y=ratio, fill=Province, label =
ratio)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(x = "Province", y = "ratio") +
  ggtitle("Ratio of Population Who Born Abroad as of 2017") +
  geom_text(size = 2.5, position = position_stack(vjust = 1.035)) +
  theme(legend.position = "none", axis.text.x = element_text(angle = 0.0,
vjust = 0.0, hjust = 0.0, size = 1))
```



Gümüshane and Bursa are surprising cities to have in top 2. Also, it's clear that ratio of people who born abroad is very low among cities in East and Southeast part of the Turkey as expected.

Part 3 - TIME Education Ranking Statistics

I used the RDS file which consist of statistics from last 3 years that one of my classmate shared with everyone. Although data gathering part is done, it does not mean we can start analysis immediately. There are some steps to clean and shape the data.

```
#Read and Combine the Data & Add Year Info to Each
r_2017 <-
readRDS("C:\\Users\\kalenderoglu\\Desktop\\BDA_assignments\\ranking2017.rds")
r_2017 <- r_2017 %>%
```

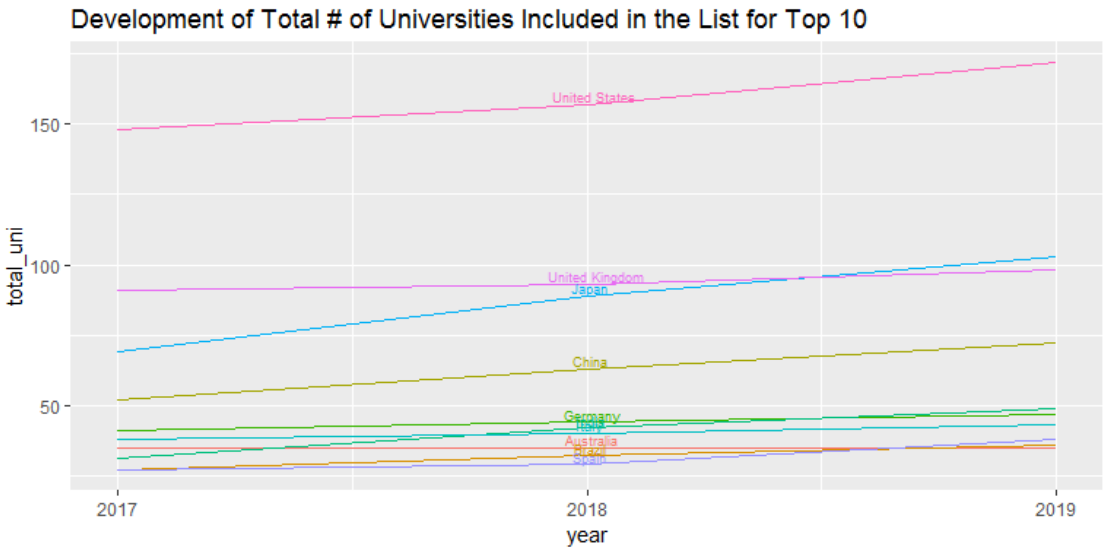
```
  mutate(year = as.numeric("2017"))
r_2018 <-
readRDS("C:\\Users\\kalenderoglu\\Desktop\\BDA_assignments\\ranking2018.rds")
r_2018 <- r_2018 %>%
  mutate(year = as.numeric("2018"))
r_2019 <-
readRDS("C:\\Users\\kalenderoglu\\Desktop\\BDA_assignments\\ranking2019.rds")
r_2019 <- r_2019 %>%
  mutate(year = as.numeric("2019"))
r_13 <- rbind(r_2017, r_2018, r_2019)
```

All variables except the year is in factor format, so we need to change type of some variables to relevant formats.

```
r_13[,3] <- as.character(r_13[,3])
r_13[,20] <- as.character(r_13[,20])
r_13[,1] <- as.numeric(as.character(r_13[,1]))
r_13[,2] <- as.numeric(as.character(r_13[,2]))
r_13[,4] <- as.numeric(as.character(r_13[,4]))
r_13[,6] <- as.numeric(as.character(r_13[,6]))
r_13[,12] <- as.numeric(as.character(r_13[,12]))
r_13[,14] <- as.numeric(as.character(r_13[,14]))
r_13[,22] <- as.numeric(as.character(r_13[,22]))
r_13[,5] <- as.numeric(gsub("=", "", as.character(r_13[,5])))
r_13[,23] <- as.numeric(sub("%", "", as.character(r_13[,23])))
r_13[,24] <- as.numeric(substr((as.character(r_13[,24])), 1, 2))
r_13[,24] <- ifelse(is.na(r_13[,24]), mean(r_13[,24], na.rm = TRUE),
r_13[,24])
```

First, i want to get picture on which countries dominate the list and how that evolved during the last 3 years

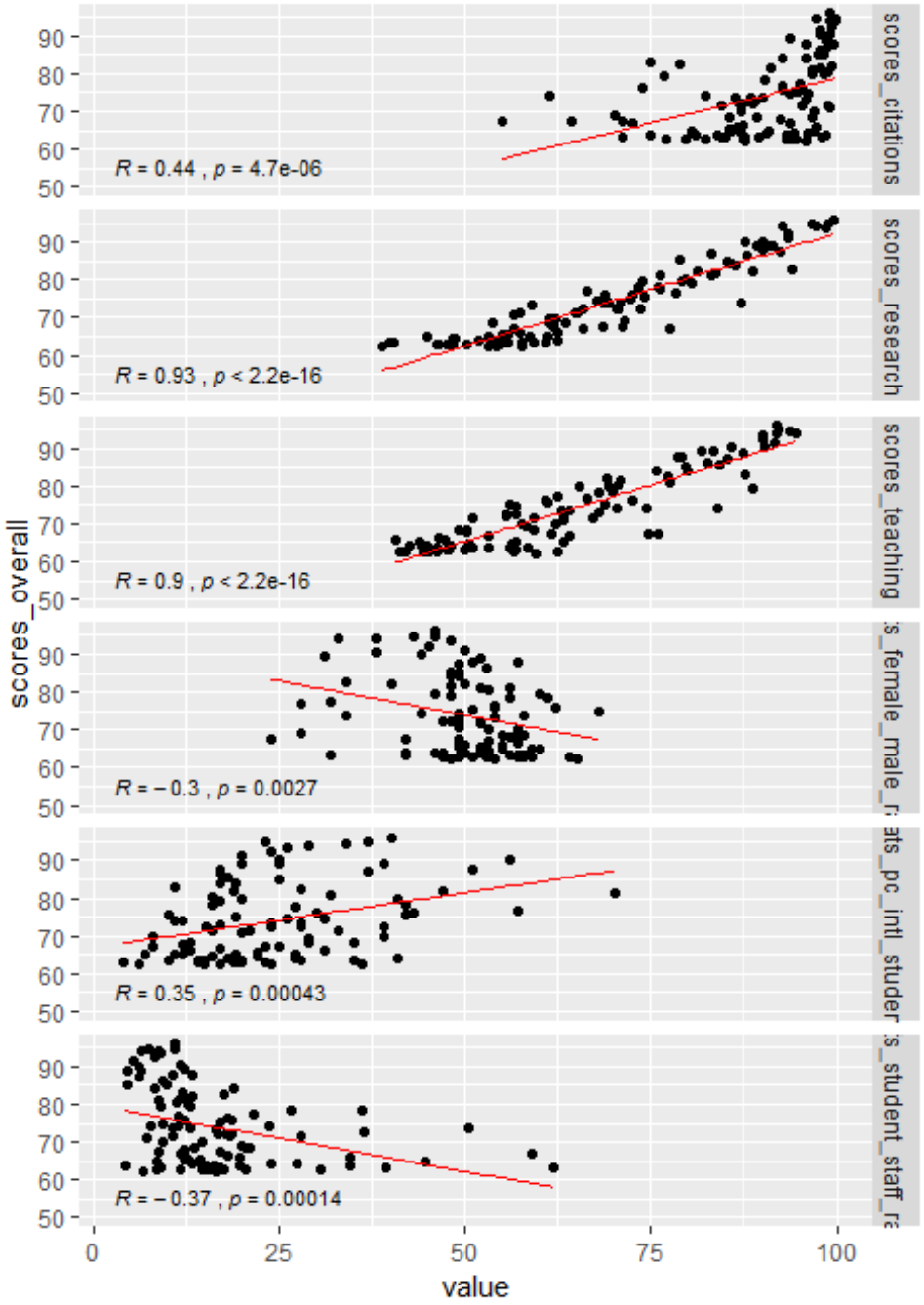
```
#Select Top 10 Ranked Country in 2019 and Count the Total Universities
r_13_v2 <- r_13 %>%
  select(year, location) %>%
  group_by(year, location) %>%
  mutate(no_of_uni = n()) %>%
  summarize(total_uni = mean(no_of_uni)) %>%
  filter(location %in% c("United States", "Japan", "United Kingdom", "China",
"India", "Germany", "Italy", "Spain", "Brazil", "Australia")) %>%
  arrange(desc(total_uni))
#Plotting
r_13_v2 %>% ggplot(aes(x= year, y = total_uni, col = location)) +
  geom_line() +
  theme(legend.position="none") +
  scale_x_continuous(breaks = c(2017, 2018, 2019)) +
  ggtitle("Development of Total # of Universities Included in the List for
Top 10") +
  geom_text(data = r_13_v2[r_13_v2$year == "2018",], aes(label = location),
hjust = .4, vjust = -.2, size = 2.4)
```



United States dominates the list in terms of number of universities by far but it's clear that some countries are on the striking rise. Japan has passed the United Kingdom in 2019. India has passed the Germany in 2019. China grows at a quiet high pace following Japan's attack. Lastly, United Kingdom can be classified as a "stable country" compared to the trend of other top locations. Although United States is still the dominant leader in the domain, we can conclude that East countries are on the way to catch and even leave behind the European countries. This can be associated with the development of the economic strength in those countries and maybe the recognition of the importance of education to compete globally in such a technology driven world might triggered East countries to invest more in high education.

As a last step, it would be insightful to pick some metrics and investigate correlation between these metrics and overall score by countries. Scatter plots would help us on that. Let's prepare the relevant data frame to feed into the graph:

```
#Data Preparation
r_13_v3 <- r_13 %>%
          filter(year == "2019") %>%
          arrange(scores_overall_rank) %>%
          head(100) %>%
          select(name, scores_overall, scores_citations, scores_research,
scores_teaching, stats_student_staff_ratio, stats_pc_intl_students,
stats_female_male_ratio) %>%
          tbl_df() %>%
          gather(metric, value, -name, -scores_overall)
#Plotting
r_13_v3 %>%
  ggplot(aes(x=value, y = scores_overall)) +
  geom_jitter() +
  facet_grid(metric ~ .) +
  geom_smooth(method = lm, color = "red", size = .3, alpha = 0) +
  stat_cor(method = "pearson", label.x = 3, label.y = 55, size=3)
```



Research score and teaching score are highly correlated with overall score as expected but it seems that research has more weight on overall score as highest correlation is observed for it. It's surprising for me to have low correlation for citation. It can be explained that since quality is more important than quantity in citation perspective, maybe it's logical that high citation score does not yield high score/rank all the time.
Although the correlation coefficients are low for remaining variables, it's worth to mention that increasing student_staff_ratio make it a bit more difficult to have higher overall score as collaboration & close relationship opportunity between student and university staff becomes wekaer.