Assignment 1

Code

Subhash Bharadwaj Pemmaraju, Raghu Mohan Sanugommula

30th January 2019

Introduction: Initial Data Analysis

The data for this analysis consists of universities from 5 states [1] -

- California (CA)
- Massachusetts (MA)
- New York (NY)
- North Carolina (NC)
- Washington (WA)

An Initial count of universities is made by State and by State and by State and Sector. The sectors have been renamed from coded numbers to descriptions using **factor** command [2]

Hide

```
sum1%>%kable("html")%>%kable_styling(bootstrap_options="striped",full_width=FALSE)
```

STABBR	Number_of_Universities
CA	100
MA	68
NC	55
NY	133
WA	19

Hide

```
sum2%>%kable("html")%>%kable_styling(bootstrap_options="striped",full_width=FALSE)
```

STABBR	SECTOR.f	Number_of_Universities
CA	Public, >=4 yrs	32
CA	Non-Profit, >=4 yrs	60
CA	Profit, >=4 yrs	7
CA	Profit, 2 yrs	1
MA	Public, >=4 yrs	13

STABBR	SECTOR.f	Number_of_Universities
MA	Non-Profit, >=4 yrs	54
MA	Profit, >=4 yrs	1
NC	Public, >=4 yrs	16
NC	Non-Profit, >=4 yrs	38
NC	Non-Profit, 2 yrs	1
NY	Public, >=4 yrs	39
NY	Non-Profit, >=4 yrs	87
NY	Profit, >=4 yrs	3
NY	Non-Profit, 2 yrs	4
WA	Public, >=4 yrs	7
WA	Non-Profit, >=4 yrs	11
WA	Profit, >=4 yrs	1

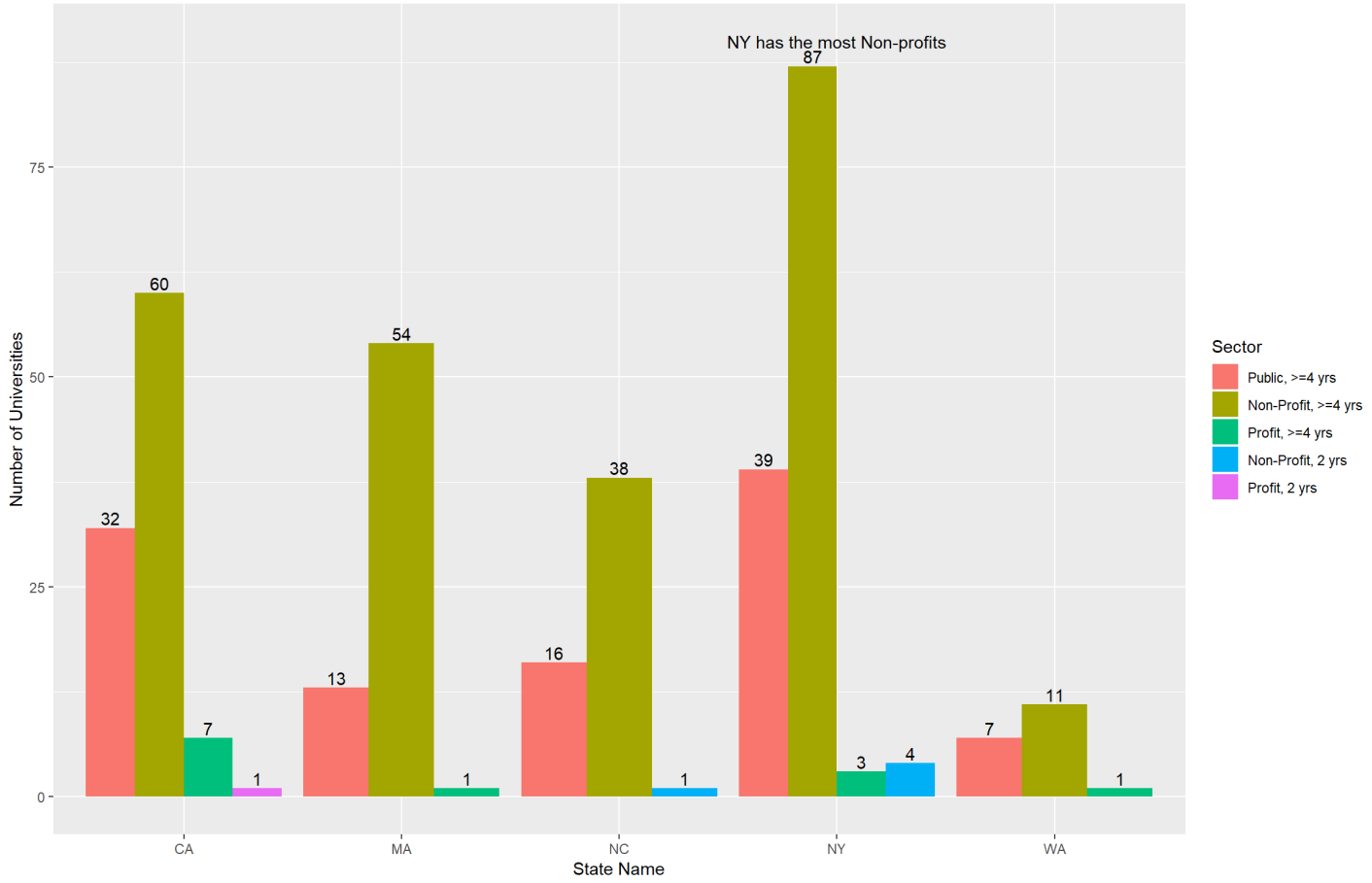
The formatting of the tables is achieved by using **kable styling** [3]

As can be seen from the tables, there are is a distinct difference between states on how many universities they have and in different sectors. A **Bar Plot** of the data is made and reveals that New York has the most number of non-profits (87) as can be seen below[4],[5],[6]:

Hide

```
ggplot(sum2, aes(x=STABBR, y=Number_of_Universities, fill=SECTOR.f))+
  geom_bar(aes(fill=SECTOR.f), position = "dodge", stat="identity")+
  ggtitle("Chart 1: Number of universities by state and category")+
  xlab("State Name")+ylab("Number of Universities")+
  geom_text(aes(label = sum2$Number_of_Universities), position=position_dodge(width=0.9), vjust=
-0.25)+
  labs(fill="Sector")+
  annotate("text", x = 4, y = 90, label = "NY has the most Non-profits")
```

Chart 1: Number of universities by state and category



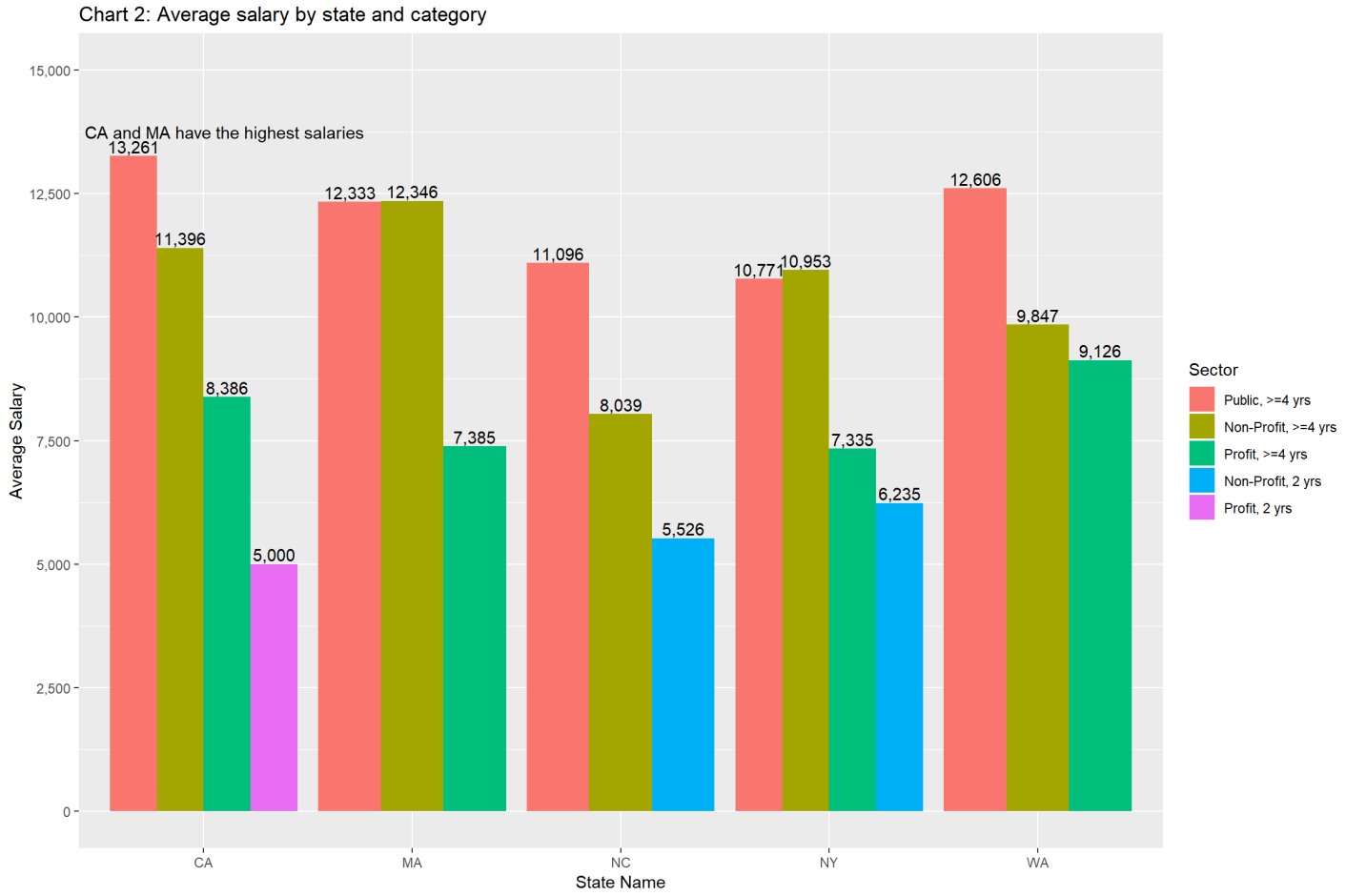
Analysis of Salaries

Salary Distribution

A chart of the average salaries by state and sector reveals surprisingly that on average public and non-profits with 4 year programs or above have higher salaries than for-profit programs. Furthermore we find that CA and MA in general have the highest salaries for these sectors [7].

Hide

```
ggplot(sum3, aes(x=STABBR, y=Average_Salaries, fill=SECTOR.f))+
  geom_bar(aes(fill=SECTOR.f), position = "dodge", stat="identity")+
  ggtitle("Chart 2: Average salary by state and category")+
  xlab("State Name")+ylab("Average Salary")+
  geom_text(aes(label = scales::comma(sum3$Average_Salaries)), position=position_dodge(width=0.9
), vjust=-0.25)+
  labs(fill="Sector")+
  annotate("text", x = 1.1, y = 13750, label = "CA and MA have the highest salaries")+
  scale_y_continuous(breaks = seq(0,15000, by=2500), limits = c(0,15000), labels = comma)
```



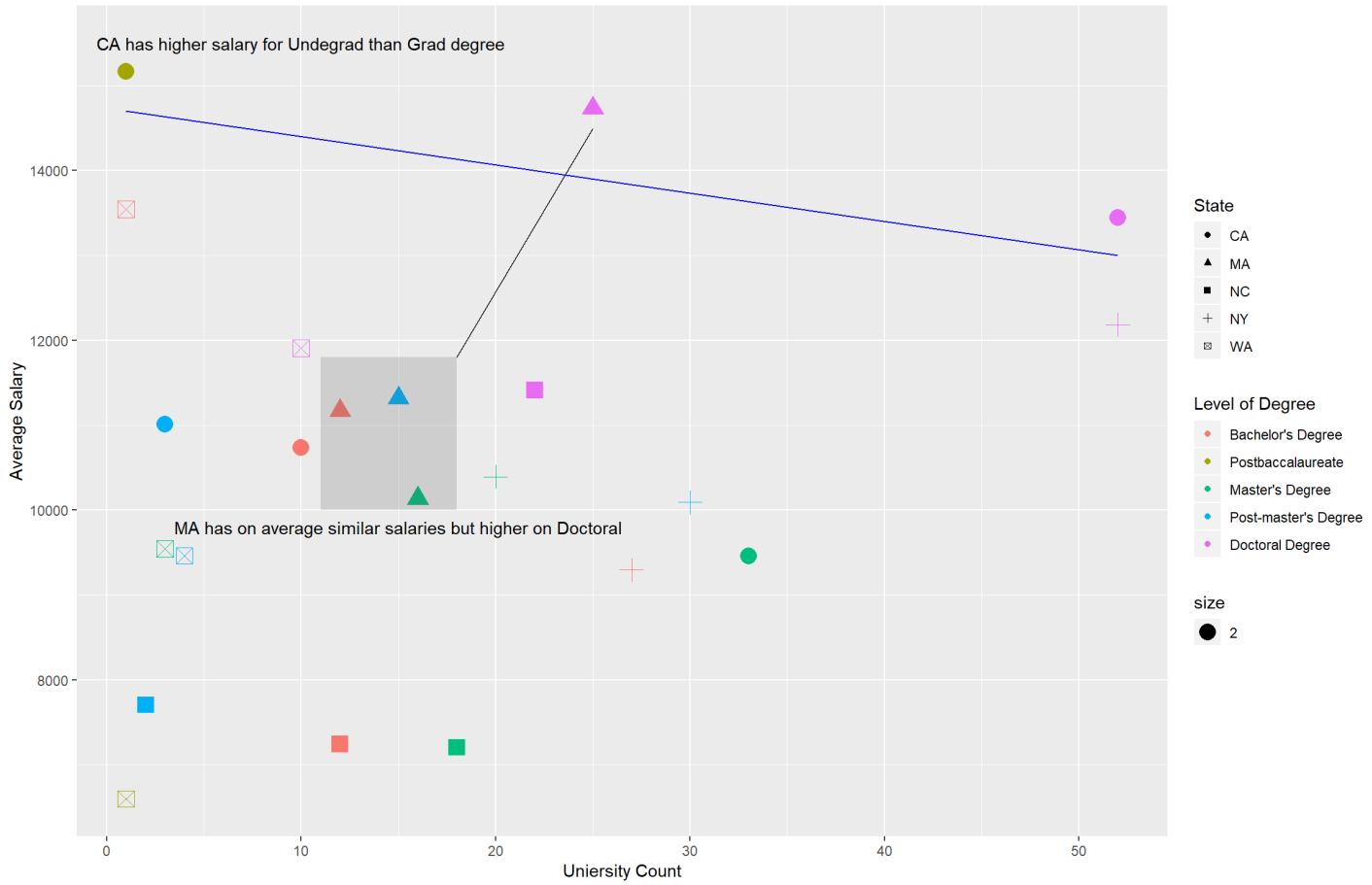
Further Analysis of Salary Distribution

Further analysis of the salary data is carried out by plotting average salary against university count taking into account the state and level of degrees offered. As can be seen in **Chart 3**, universities in California offer on average, higher salary for undergraduate programs than universities that offer PhD programs. However, the salary of universities in Massachusetts is more closely aligned across different levels of programs offered. However, MA does offer a much higher salary for universities that offer PhD programs - the highest among all 5 states [8],[9].

Hide

```
ggplot(sum4, aes(x=Number_of_Universities, y=Average_Salaries, group=STABBR, fill=HLOFFER.f))+
  geom_point(aes(shape=STABBR, color=HLOFFER.f, size=2))+
  ggtitle("Chart 3: Average salary vs University count")+
  xlab("Uniersity Count")+ylab("Average Salary")+
  labs(fill="Level of Degree", color = "Level of Degree")+
  labs(shape="State")+
  annotate("segment", x=1, y=14700, xend=52, yend=13000, color="blue")+
  annotate("text", x =10, y=15500, label="CA has higher salary for Undegrad than Grad degree")+
  annotate("rect", xmin=11, xmax=18, ymin=10000,ymax=11800, alpha=0.2)+
  annotate("segment", x=18,y=11800,xend=25,yend=14500,color="black")+
  annotate("text", x=15, y=9800, label = "MA has on average similar salaries but higher on Docto
ral")
```

Chart 3: Average salary vs University count



References

1. How to use or/and in dplyr to subset a data.frame. Stackoverflow. Link: https://stackoverflow.com/questions/24319747/how-to-use-or-and-in-dplyr-to-subset-a-data-frame (https://stackoverflow.com/questions/24319747/how-to-use-or-and-in-dplyr-to-subset-a-data-frame)↩
2. FACTOR VARIABLES | R LEARNING MODULES. Institute for Digital Research and Education, UCLA. Link: https://stats.idre.ucla.edu/r/modules/factor-variables/ (https://stats.idre.ucla.edu/r/modules/factor-variables/)↩
3. Zhu, Hao. (2019). Create Awesome HTML Table with knitr::kable and kableExtra. GitHub. Link: https://haozhu233.github.io/kableExtra/awesome_table_in_html.html#overview (https://haozhu233.github.io/kableExtra/awesome_table_in_html.html#overview)↩
4. ggplot2 legend : Easy steps to change the position and the appearance of a graph legend in R software. Statistical tools for high-throughput data analysis. Link: http://www.sthda.com/english/wiki/ggplot2-legend-easy-steps-to-change-the-position-and-the-appearance-of-a-graph-legend-in-r-software (http://www.sthda.com/english/wiki/ggplot2-legend-easy-steps-to-change-the-position-and-the-appearance-of-a-graph-legend-in-r-software)↩
5. How to change legend title in ggplot. Stackoverflow. Link: https://stackoverflow.com/questions/14622421/how-to-change-legend-title-in-ggplot (https://stackoverflow.com/questions/14622421/how-to-change-legend-title-in-ggplot)↩

6. How to put labels over geom_bar for each bar in R with ggplot2. Stackoverflow. Link: https://stackoverflow.com/questions/12018499/how-to-put-labels-over-geom-bar-for-each-bar-in-r-with-ggplot2 (https://stackoverflow.com/questions/12018499/how-to-put-labels-over-geom-bar-for-each-bar-in-r-with-ggplot2)↩

7. Formatting ggplot2 axis labels with commas (and K? MM?) if I already have a y-scale. Stackoverflow. Link: https://stackoverflow.com/questions/37713351/formatting-ggplot2-axis-labels-with-commas-and-k-mm-if-i-already-have-a-y-sc (https://stackoverflow.com/questions/37713351/formatting-ggplot2-axis-labels-with-commas-and-k-mm-if-i-already-have-a-y-sc)↩

8. ggplot2 point shapes. Statistical tools for high-throughput data analysis. Link: http://www.sthda.com/english/wiki/ggplot2-point-shapes (http://www.sthda.com/english/wiki/ggplot2-point-shapes)↩

9. Bhaskar VK. hrbrthemes : Additional Themes and Theme Components for ‘ggplot2’. GitHub. Link: https://bhaskarvk.github.io/hrbrthemes/ (https://bhaskarvk.github.io/hrbrthemes/)↩