Project Final Report

Members

- Hyun Ko
- **Kyle Yeo**
- Riley Larget
- Winsten Coellins

Introduction

This analysis was chosen in order to point out several questions that students, especially international students studying abroad in the United States, have been asking such as "Does the type of school that I choose have an impact in my income after I graduate?" and "Does the region of school really matter when it comes to the level of income at a workplace?". Furthermore, students always seem to be interested in the topics like which types of college tends to have the alumni with the highest average salary. Therefore, in this analysis, we are going to focus on questions that could answer the concerns of international students and even domestic students by asking the relation of college type to salary, the region of college to salary, and the initial salary to mid-career salary. We found that there were statistically significant differences in the level of salary depending on the types of college and there are no significant differences in the level of salary depending on the location of school. Therefore, the type of college is more important than region. We also found that the salary increases faster over time for lower starting salaries for all types of college, with the exception of Ivy League colleges.

Questions of Interest:

These are the questions of interest that we will analyze.

(a) Is the type of college related to salary?

(b) Is the type of region related to salary?

(c) Which undergraduate major has the highest starting median salary?

(d) Which undergraduate major has the highest mid career salary?

(e) Which type of school has the highest starting median salary?

(f) Which type of school has the highest mid career median salary?

(g) Which region of schools has the highest starting median salary?

(h) Which region of school has the highest mid career median salary?

(i) Between the type of college and type of region, which one is more important to salary?

(j) How does the initial salary relate to the increase in salary over time?

Background

About the raw data

There are three datasets, degrees-that-pay-back.csv, salaries-by-college-type.csv, and salaries-by-region.csv. we got this data from Kaggle, which is the official website that offers tons of different datasets to public. Each dataset was gathered from reports made to PayScale Inc. The first data called degrees-that-pay-back.csv consists of undergraduate majors with starting median salary and mid career median salary by stages from 10th percentile to 90th percentile. The second data called salaries-by-college-type.csv consists of the variables, school name, school type, starting median salary, and mid career salary also by stages from 10th percentile to 90th percentile. The third data also consists of similar structures with the first and second datasets with the variables, school name, region, starting median salary, and mid career median salary by stages from 10th percentile to 90th percentile. The topic of this analysis might be familiar to most of people, if not all, so that it might be not difficult to know the meaning of each variables. The variable undergraduate major indicates just the majors that are offered by college/university. The variable school name indicates the names of U.S. college or university and the region represents the location of college/universities like California. The variable school type means the type of school such as engineering school. The variable starting median salary indicates the median of the base salary and mid career median salary represents the level of salary after working for certain period of years. Lastly, the variable mid career x percentile means the level of a wage below a certain percent of workers. In the analysis, we combined the second and the third data for the research and analyze the first data separately.

Datasets and variables used in the anaylsis

As we mentioned above, there are a number of variables in each datasets, but not all of them used in the analysis. we didn't use the variables of mid career X percentile since we could well address the questions of interest without those variables.

Shortcomings of the datasets

There are few shortcomings of our datasets. First of all, the data, degrees-that-pay-back, only has the names of the undergraduate majors, not with the schools so that we couldn't combine it with the other two datasets. In addition, the variable of mid career median salary didn't specify what "mid career" means so there's no way to know about from which period of working years should be considered as mid career.

The source of the data:

- https://www.kaggle.com/wsj/college-salaries

For the rest of the report we are going to tidy up three datasets appropriately and analyze our research questions with the numerical tibbles, graphs, linear models and t-tests.

Analysis

Cleaning up the data

As we mentioned in the background, we want to clean up the datasets into an appropriate format for the analysis.

We apply two functions to the raw data. Firstly, we use a reformatting function since the original data has salaries as strings like "$79,000.00" and the function converted those into numbers like 79000. Also, we use a renaming function in order to make column names R-compliant. We then join the type dataset and region dataset by school name into a dataset of schools with both fields.

Numerical analysis

> Firstly, we want to address which type of school has the highest starting median salary and mid career salary.

```
## # A tibble: 5 x 5
##   type   mean_of_starting_~ mean_of_mid_caree~ sd_of_starting_~ sd_of_mid_caree~
##   <chr>               <dbl>              <dbl>            <dbl>            <dbl>
## 1 Ivy L~              60475             120125             3219.            3219.
## 2 Engin~              59411.            105128.            7913.            7913.
## 3 Liber~              45747.             89379.            4369.            4369.
## 4 Party               45715              84685             3686.            3686.
## 5 State               44126.             78567.            4269.            4269.
```

We made the numerical tables with the data of type_and_region, which is the combined data and calculated mean of the starting median salary and mean of the mid career median salary by the variable of school type. Also, we calculated standard deviation of both starting median salary and mid career median salary to measures the spread of a data distribution. Given the tables, we can conclude that people who graduated from Ivy League schools tend to have the highest starting median salary and mid career median salary. Just by looking at the values of standard deviation, we can say that in both cases, the data distribution of Engineering schools tends to be more spread out. We will analyze this in more detail and with the different approach in the later part.

> we also want to address which region of schools has the highest starting median salary and which region of school has the highest mid career median salary with the same approach as we did above.

```
## # A tibble: 5 x 5
##   region  mean_of_starting_~ mean_of_mid_care~ sd_of_starting_~ sd_of_mid_caree~
##   <chr>                <dbl>             <dbl>            <dbl>            <dbl>
## 1 Califo~              50156.            91633.           8247.            8247.
## 2 Northe~              49267.            93519.           7796.            7796.
## 3 Southe~              44288.            80303.           4322.            4322.
## 4 Western              44151.            78136.           3926.            3926.
## 5 Midwes~              43802.            77638.           4580.            4580.
```

We made the numerical tables with the data of type_and_region, which is the combined data and calculated mean of the starting median salary and mean of the mid career median salary by the variable of region at this time. Also, we calculated standard deviation of both starting median salary and mid career median salary to measures the spread of a data distribution. Given the tables, we can conclude that people who graduated from schools located in the region of California tend to have the highest starting median salary, while the people who graduated from schools located in Northeastern tend to have the highest mid career median salary and we conclude that this is because all the Ivy League schools are located in the region of Northeastern so that Ivy League schools have a great impact on mid career salary. Just by looking at the values of standard deviation, we can say that the data distribution of starting median salary of California tends to be more spread out and the data distribution of mid career salary of Northeastern tends to be more spread out. We will analyze this in more detail and with the different approach in the later part.

> Next, we will analyze which undergraduate major is the best for having the highest starting median salary and mid career salary.

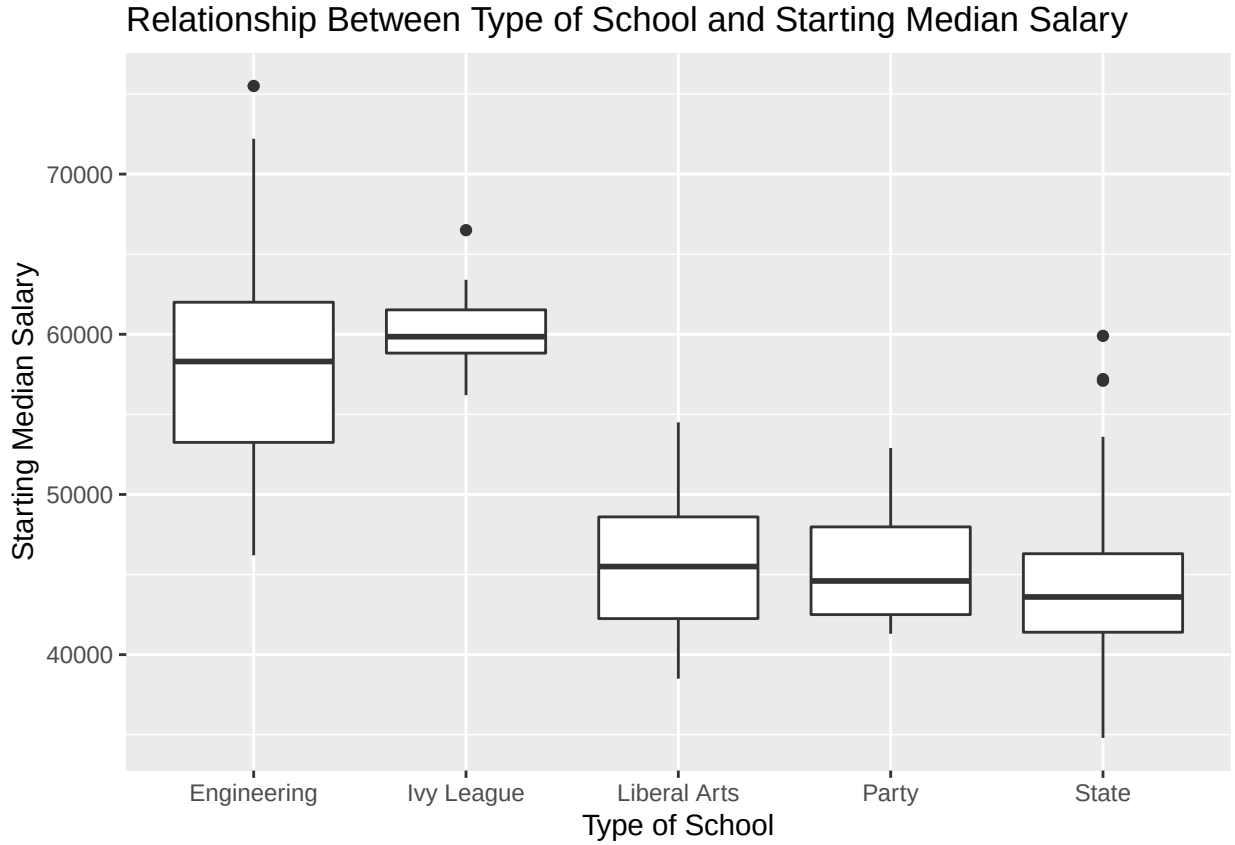
```
## # A tibble: 50 x 3
##    undergraduate_major    mean_starting mean_mid_career
##    <chr>                          <dbl>           <dbl>
##  1 Physician Assistant            74300           91700
##  2 Chemical Engineering           63200          107000
##  3 Computer Engineering           61400          105000
##  4 Electrical Engineering         60900          103000
##  5 Mechanical Engineering         57900           93600
##  6 Aerospace Engineering          57700          101000
##  7 Industrial Engineering         57700           94700
##  8 Computer Science               55900           95500
##  9 Nursing                        54200           67000
```

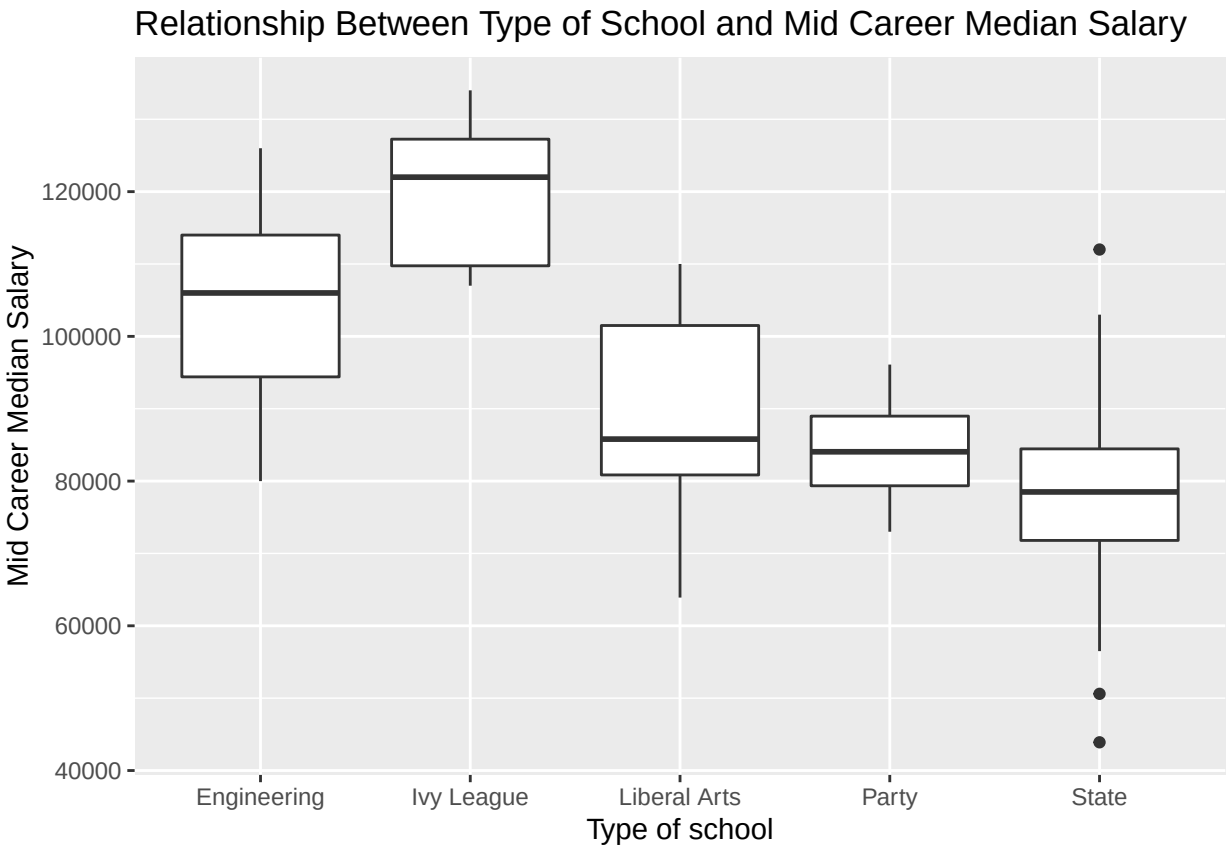
```
## 10 Civil Engineering              53900           90500
## # ... with 40 more rows
```

We made the above numerical table in order to analyze the relationship between the undergraduate majors and starting and mid career salary. According to the tables, we can conclude that people who are graduated with the major of physician assistant will be more likely to have a high starting salary and the people who are graduated with the major of chemical engineering will be more likely to have a high mid career salary.

Graphical analysis

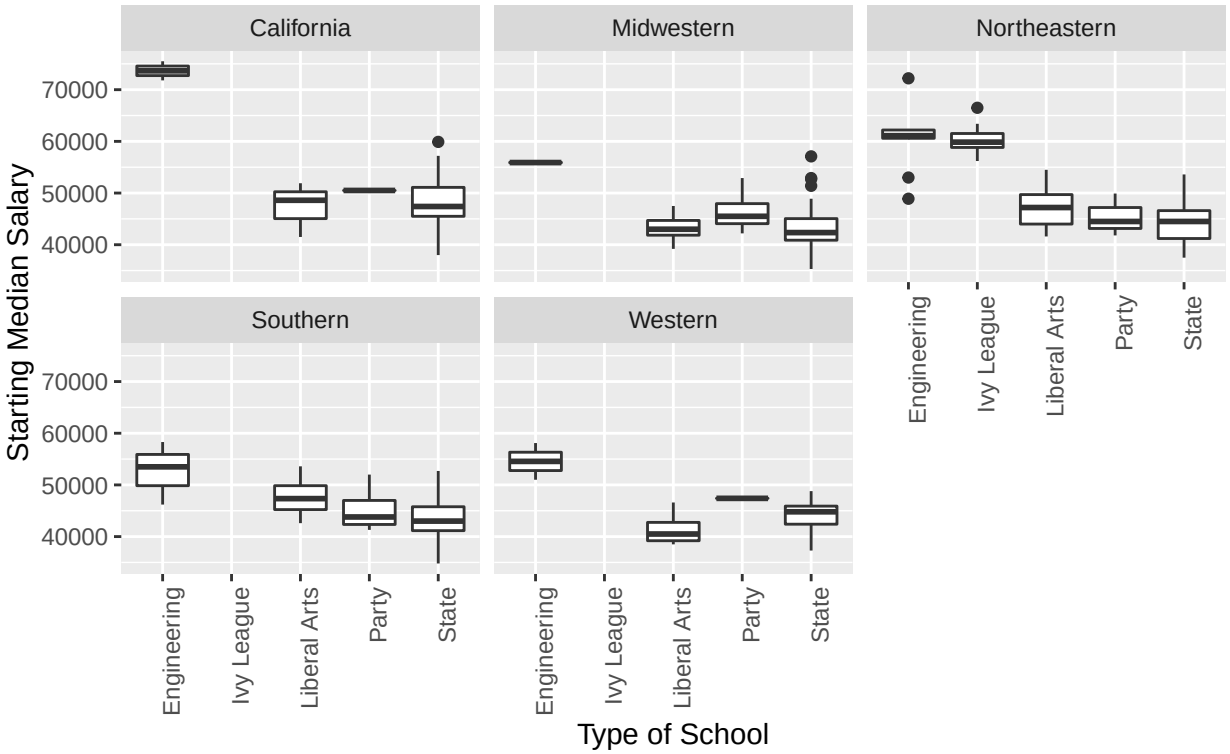
For this part, we will explore the relationship between the type of school and starting, mid career median salary with the visualization by making box plots.

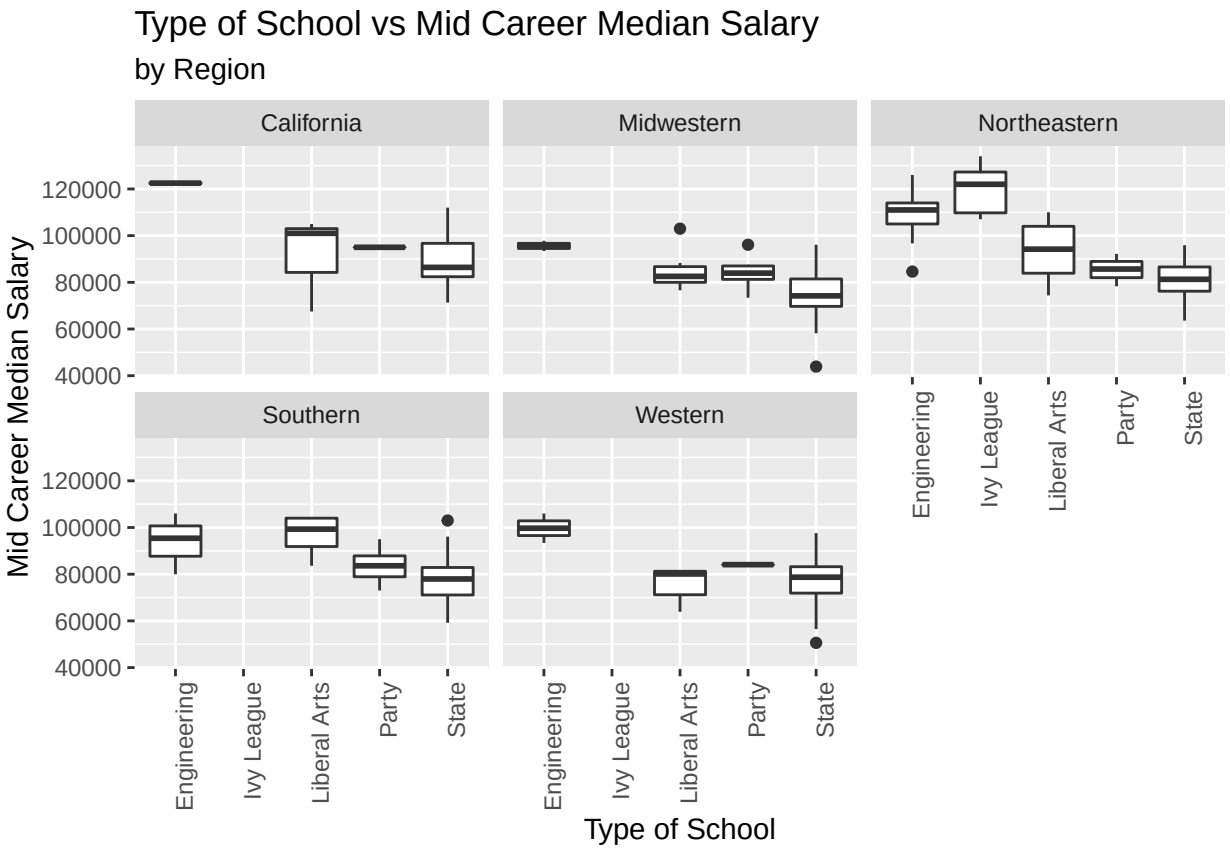




According to the both plots, we can more easily conclude that people who are graduated from Ivy league schools tend to have the highest both starting median salary and mid career median salary. However, by virtue of the characteristic of the boxplot, we could figure out that there are some outliers in each type of schools. Furthermore, each types of school has different salaries so that we can say that the type of school related to the level of individual's income.

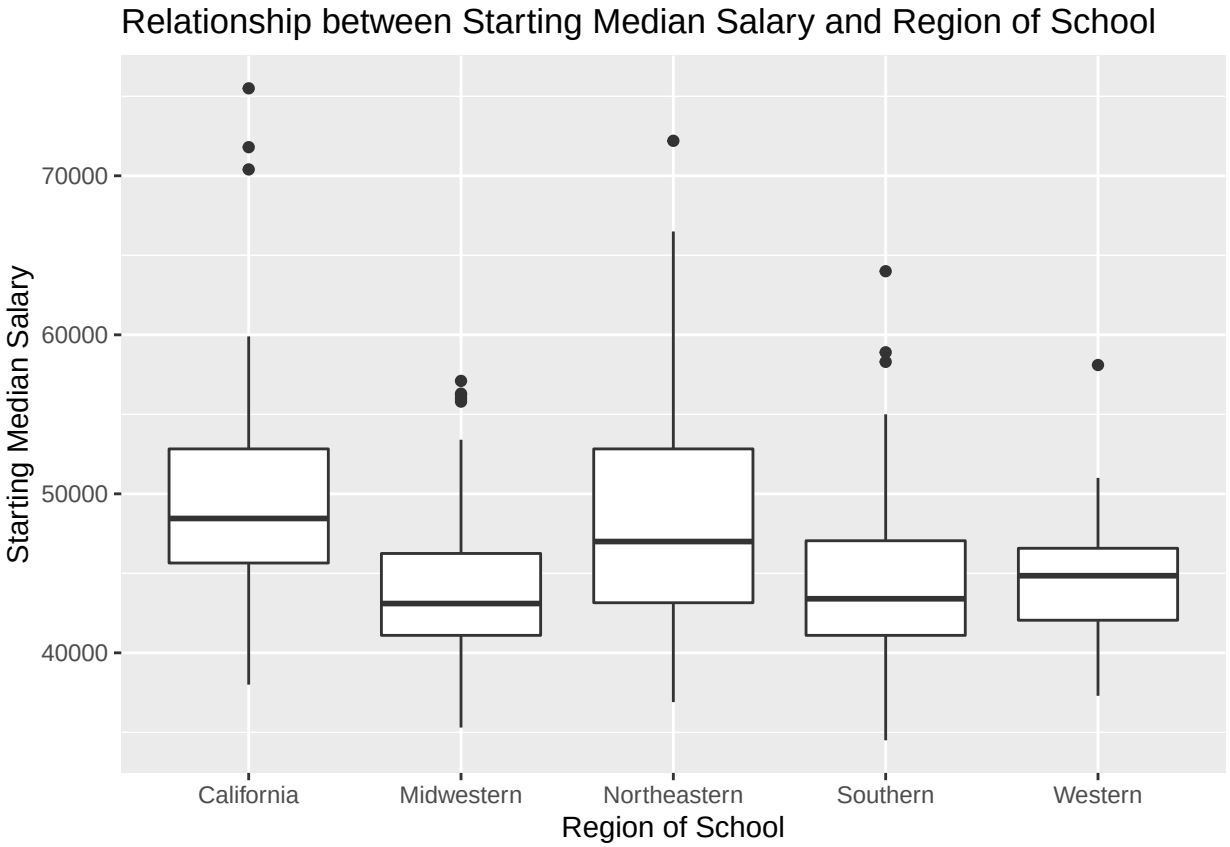
> For this part, we will also explore the relationship between the type of school and starting, mid career median salary with the visualization by making box plots. However, at this time, we will use the combined dataset and separate the types of school depends on the region of the school.

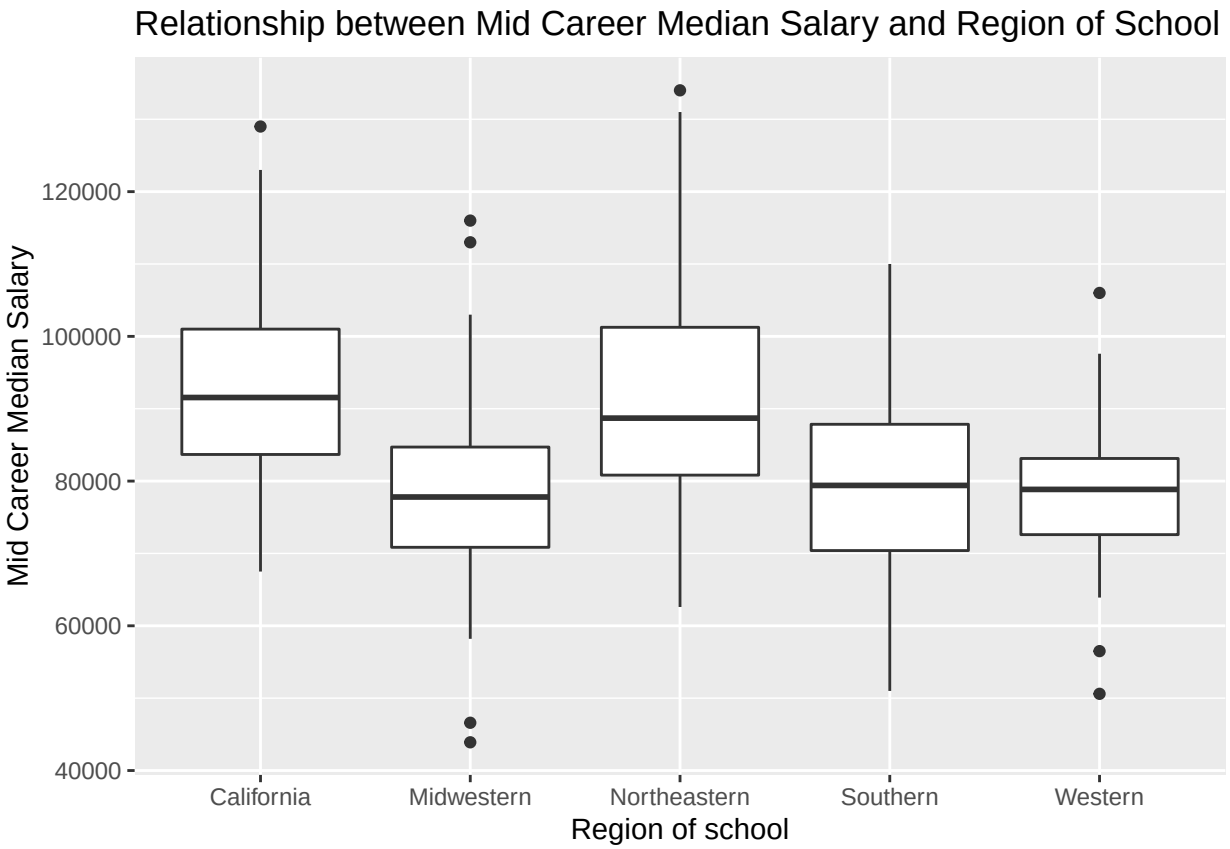

Type of School vs Starting Median Salary
by Region
California
Midwestern
Northeastern
Southern
Western
Starting Median Salary
70000
60000
50000
40000
Engineering
Ivy League
Liberal Arts
Party
State
Type of School



According to the graphs, we expect to figure out the relationship between the type of school and starting, mid career salary with the region of school at the same time. It's important to note that engineering schools in every region tend to have the highest median of starting salary. In case of mid career median salary, the regions of California, Midwestern, and Western do not show any significant difference from the case of starting median salary, but in the region of Northeastern and Southern, the Ivy League schools and Liberal Arts schools tend to have the highest mid career median salary, respectively.

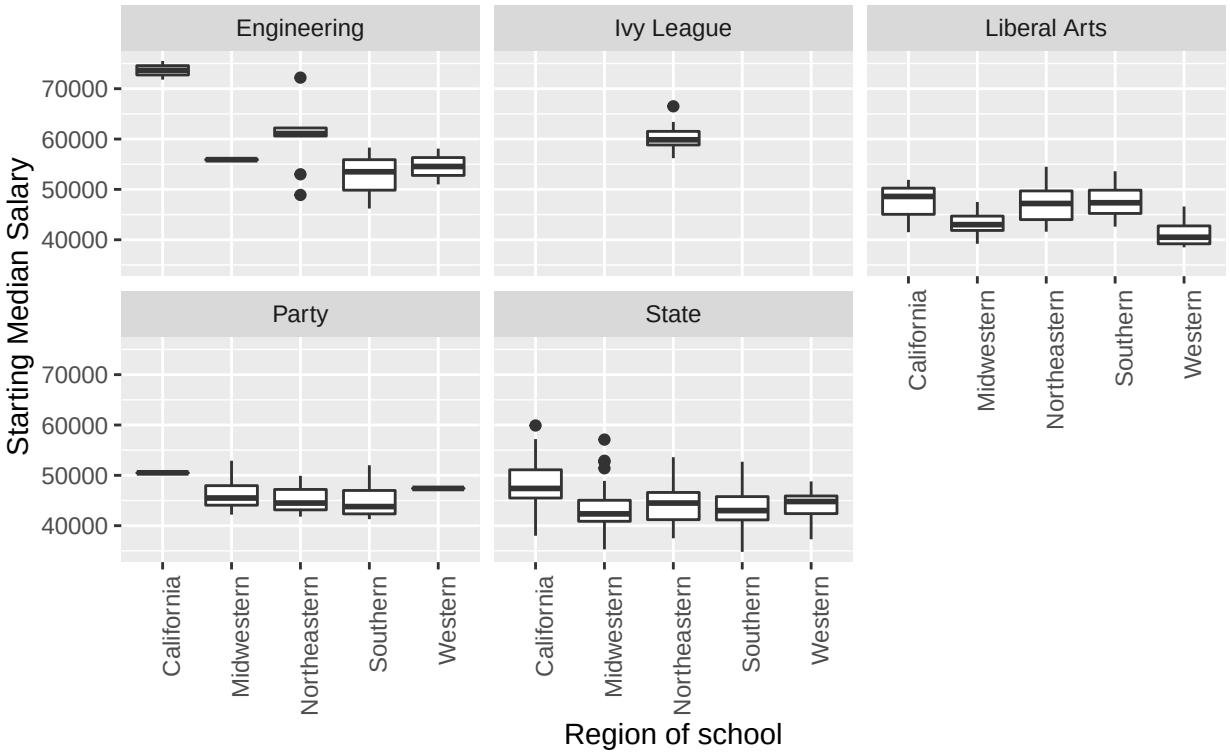
In this part, we will explore the relationship between the location of school and starting, mid career median salary with the visualization by making box plots.

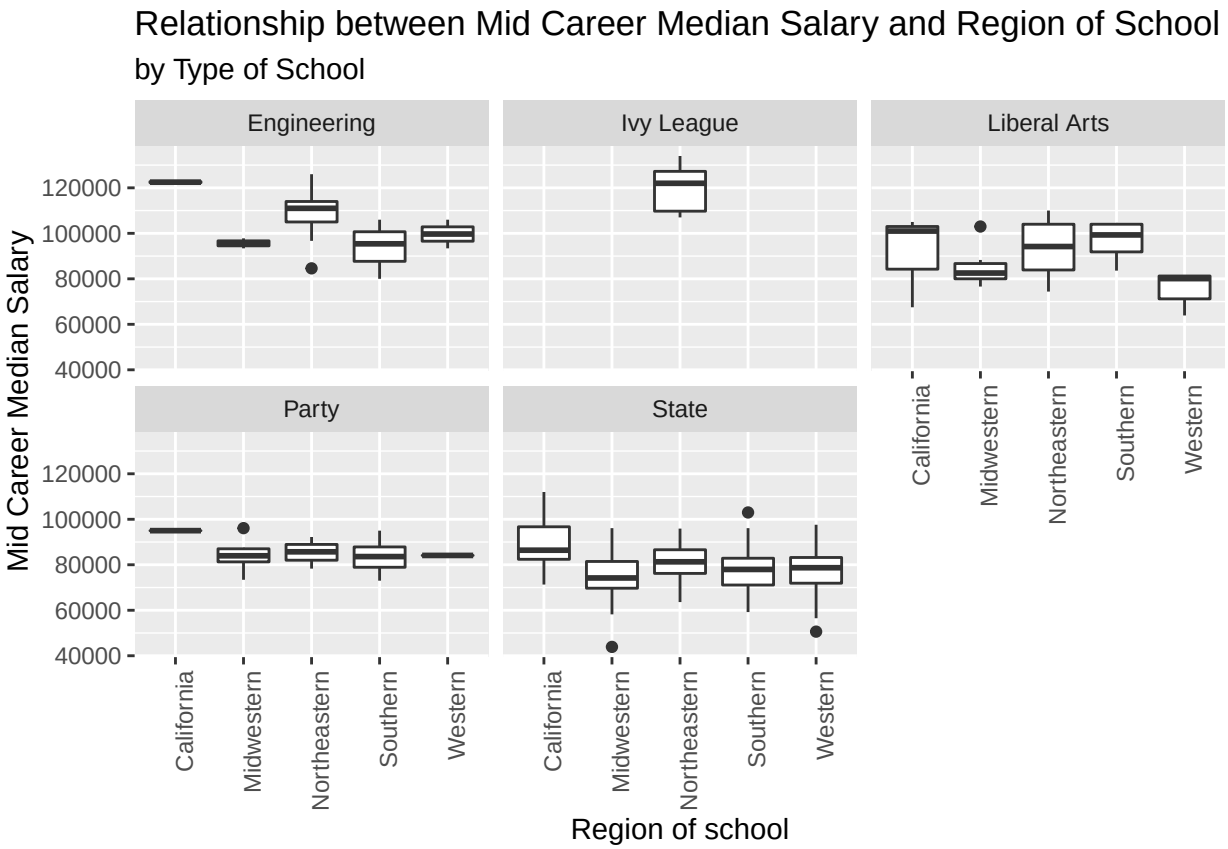

Relationship between Starting Median Salary and Region of School
Starting Median Salary
70000
60000
50000
40000
California
Midwestern
Northeastern
Southern
Western
Region of School



According to the both plots, we can conclude that people who are graduated from schools in California and Northeastern tend to have the highest both starting median salary and mid career median salary. However, it's difficult to differentiate two regions so that we can say that the numerical approach is the better way for this analysis. Also, by virtue of the characteristic of the boxplot, we could figure out that there are some outliers in each region of schools.

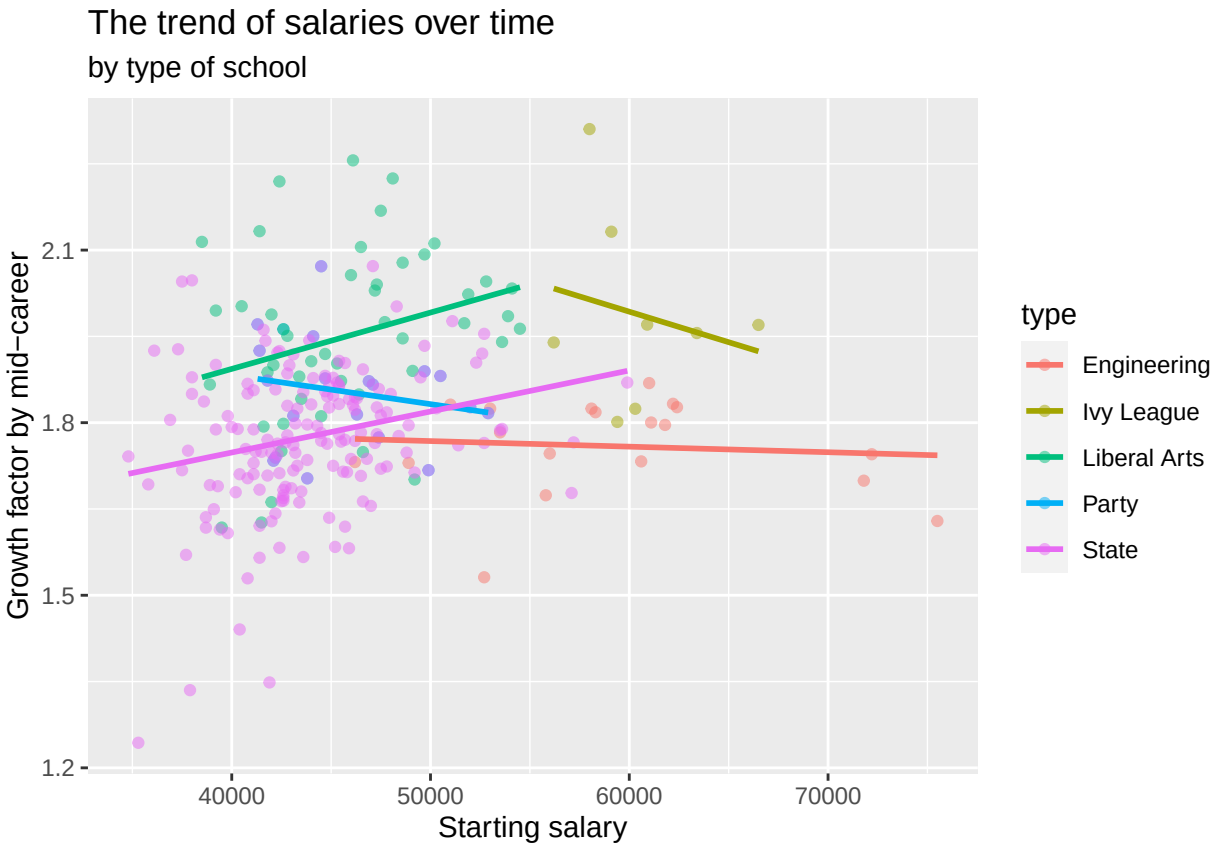
In this part, we will also explore the relationship between the region of school and starting, mid career median salary with the visualization by making box plots. However, at this time, we will use the combined dataset and separate the regions of school depends on the type of the school.

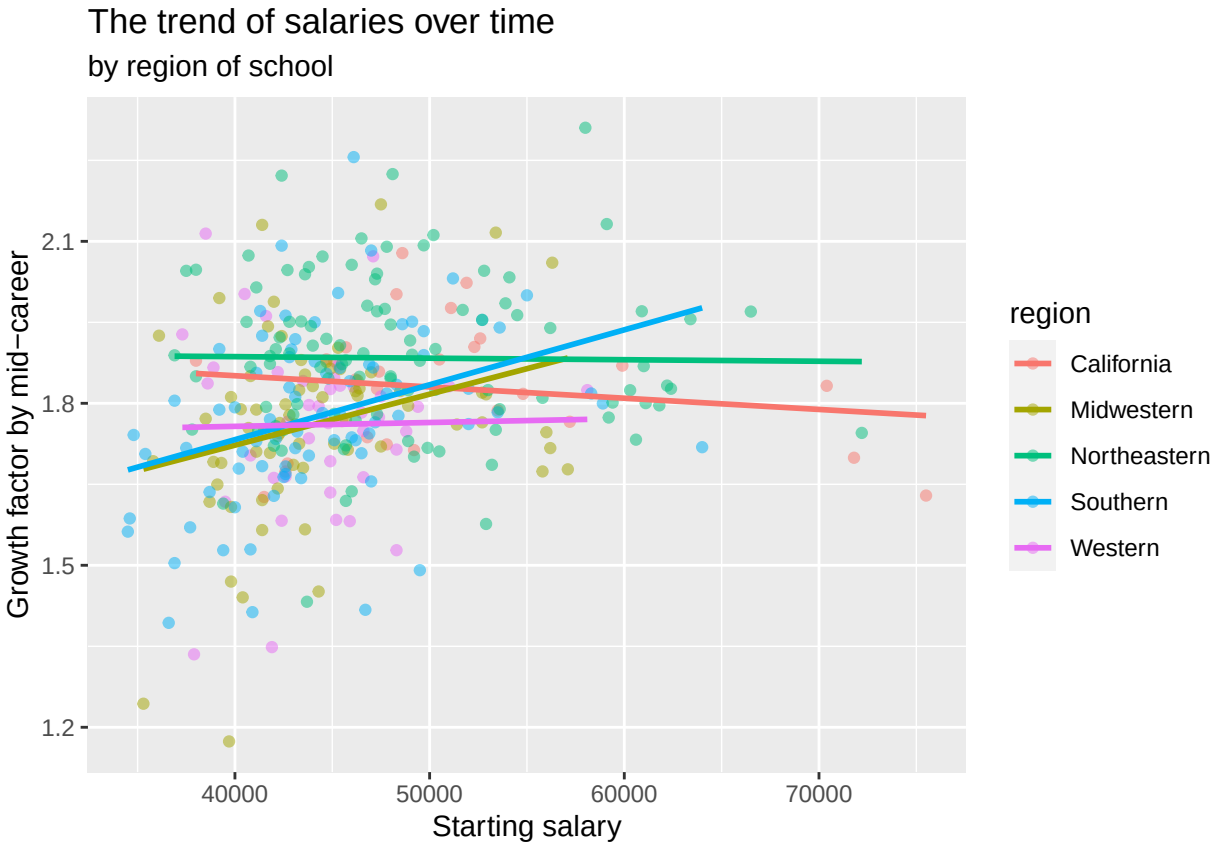

Relationship between Starting Median Salary and Region of School
by Type of School
Engineering
Ivy League
Liberal Arts
Party
State
Starting Median Salary
70000
60000
50000
40000
California
Midwestern
Northeastern
Southern
Western
Region of school



According to the both graphs, we expect to figure out the relationship between the region of school and starting, mid career salary depending on the type of school at the same time. It's important to note that regardless of the type of schools, people who are graduated from schools located in the region of California will be more likely to have better income comparing to others in the same field. However, since Ivy League schools only located in the region of Northeastern, we can say that the Ivy League schools are exeptions for this case.

Linear models of growth by starting salary





```
## # A tibble: 5 x 6
## # Groups:   type [5]
##   type        salary_dependen~ min_start max_start min_mid_career max_mid_career
##   <chr>                  <dbl>     <dbl>     <dbl>          <dbl>          <dbl>
## 1 Liberal Ar~       0.00000980     38500     54500          63900         110000
## 2 State             0.00000713     34800     59900          43900         112000
## 3 Engineering      -0.00000245     46200     75500          80000         126000
## 4 Party            -0.00000505     41300     52900          73000          96100
## 5 Ivy League       -0.0000106      56200     66500         107000         134000
```

```
## # A tibble: 5 x 6
## # Groups:   region [5]
##   region      salary_dependen~ min_start max_start min_mid_career max_mid_career
##   <chr>                  <dbl>     <dbl>     <dbl>          <dbl>          <dbl>
## 1 Southern          0.00000830      34800     58300          59200         106000
## 2 Midwestern        0.00000494      35300     57100          43900         103000
## 3 Western           0.00000261      37300     58100          50600         106000
## 4 Northeaste~       0.000000438     37500     72200          63600         134000
## 5 California       -0.00000258      38000     75500          67500         123000
```

We visualized and calculated the slope of each type and region of school with respect to starting salary and mid career salary growth rate. Positive slopes here show a widening distribution of salaries at mid career compared to at career start, while negative slopes show a tightening distribution of salaries. Relative magnitude of slopes show the strength of this effect.

Hypothesis test with p values

Comparing Engineering or Ivy League salaries to other school type salaries:
$H_0 : \mu_{Eng,Ivy} = \mu_{Lib,State,Party}$
$H_a : \mu_{Eng,Ivy} \neq \mu_{Lib,State,Party}$
For starting median salary, p-value $= 1.0851022 \times 10^{-11} < 0.01$. Therefore we can reject the null hypothesis that they have the same mean.
For mid-career median salary, p-value $= 6.4533533 \times 10^{-11} < 0.01$. Therefore we can reject the null hypothesis that they have the same mean.

Comparing Ivy League salaries to Engineering salaries:
$H_0 : \mu_{Eng} = \mu_{Ivy}$
$H_a : \mu_{Eng} \neq \mu_{Ivy}$
For the starting median salary, p-value $= 0.6307359 > 0.05$. Therefore we cannot reject the null hypothesis that they have the same mean.
For the mid-career median salary, p-value $= 0.0061883 < 0.01$. Therefore we can reject the null hypothesis that they have the same mean.

Comparing Engineering salaries to other school type salaries:
$H_0 : \mu_{Eng} = \mu_{Lib,State,Party}$
$H_a : \mu_{Eng} \neq \mu_{Lib,State,Party}$
For the starting median salary, p-value $= 3.3997664 \times 10^{-7} < 0.01$. Therefore we can reject the null hypothesis that they have the same mean.
For the mid-career median salary, p-value $= 3.6446443 \times 10^{-7} < 0.01$. Therefore we can reject the null hypothesis that they have the same mean.

Comparing between region grouping for Party schools:
$H_0 : \mu_{Party-CA,NE} = \mu_{Party-other}$
$H_a : \mu_{Party-CA,NE} \neq \mu_{Party-other}$
For the starting median salary, p-value $= 0.6283073 > 0.05$. Therefore we cannot reject the null hypothesis that they have the same mean.

For the mid-career median salary, p-value = 0.3935479 > 0.05. Therefore we cannot reject the null hypothesis that they have the same mean.

Discussion

Provide broader interpretations of your analysis and describe how to interpret your results with respect to your questions of interest & Summarize your primary conclusions and the primary evidence that supports these conclusions.

Using the numerical summaries, we can see that Ivy league schools have the highest mean of both starting median salary and mid career median salary and California has the highest starting median salary while the Northeastern region has the highest mid career median salary. From the numerical summaries with degree data, we found that Physician Assistant major has the highest starting median salary and Chemical Engineering major has the highest mid career median salary.
For starting median salary, there is a clear dichotomy separating Ivy League and Engineering schools from the other schools. Once mid-career is reached, this gap still exists but it has narrowed. Liberal Arts schools have the fastest closing rate followed by State schools. Ivy League schools begin to separate with higher salaries than Engineering. This could be caused by post-graduate education (Law and Medical School), but that is not available in the data. For salary growth, Ivy League schools have the highest growth rate of all school types, followed by Liberal Arts, State, Party and Engineering. From the best fit slopes, we see that the starting median salary is most strongly related to mid-career median salary for Liberal Arts schools and least strongly related for Ivy League schools.
There is minimal separation by region. Where this does occur, it appears to be caused by difference in type of school by region. For example, California has the highest paying Engineering schools and the Northeast has Ivy League schools.

Discuss any potential short-comings of the analysis.

(a) Some of the school presented in the data has more than one school type, therefore, it could affect the result that we have done

(b) The degree data only has undergraduate major information, not with the school name so that it cannot be used with the type and region data.

(c) There's no information about graduate school by school or major.

(d) There are some overlappings in the data. For example, there are separate engineering schools like MIT, but also other type of universities like UW-Madison have engineering major. Thus, this overlappings in the data can affect the accuracy of an analysis on 'engineering' part.

(e) Ivy-League schools are entirely located at Northeastern region, which makes it difficult to make overall analysis across the region. What if we want to look out for 'State' schools across all regions?

Potential future directions for additional work

Different methods to address the same questions

- We could do additional hypothesis testing for more groupings.

New questions

- Do the same majors from different university have effect on salary?

New data you might collect to refine your understanding

- Our current degree data only has the undergraduate majors with salary by stages. In order to address the new question, we need to collect the data that contains not only the majors, but also the school names with salary.