# Analysis 1: UNC Salaries

Breonna Corbett

February 22, 2021

# Instructions

**Overview:** For each question, show your R code that you used to answer each question in the provided chunks. When a written response is required, be sure to answer the entire question in complete sentences outside the code chunks. When figures are required, be sure to follow all requirements to receive full credit. Point values are assigned for every part of this analysis.

**Helpful:** Make sure you knit the document as you go through the assignment. Check all your results in the created PDF file.

**Submission:** Submit via an electronic document on Gradescope. Must be submitted as an PDF file generated in RStudio.

# Introduction

Universities are typically opaque, bureaucratic institutions. To be transparent to tax payers, many public schools, such as the University of North Carolina, openly report **salary information (http://www.newsobserver.com/news/databases/public-salaries/)**. In this assignment, we will analyze this information to answer pivotal questions that have endured over the course of time. The most recent salary data for UNC-Chapel Hill faculty and staff has already been downloaded in CSV format and titled *"UNC_System_Salaries Search and Report.csv"*. If you scan the spreadsheet, you will notice that Dr. Mario is not listed. People get depressed when they see that many digits after the decimal.

To answer all the questions, you will need the R package `tidyverse` to make figures and utilize `dplyr` functions.

# Data Information

Make sure the CSV data file is contained in the folder of your RMarkdown file. First, we start by using the `read_csv` function from the `readr` package found within the tidyverse. The code below executes this process by creating a tibble in your R environment named "salary".

```
salary=read_csv("UNC_System_Salaries Search and Report.csv")
```

Now, we will explore the information that is contained in this dataset. The code below provides the names of the variables contained in the dataset.

```
names(salary)
```

```
##  [1] "Name"                  "campus2"               "dept"
##  [4] "position"              "PRIMARY_WORKING_TITLE" "hiredate"
##  [7] "exempt"                "fte"                   "employed"
## [10] "statesal"              "nonstsal"              "totalsal"
## [13] "stservyr"
```

Next, we will examine the type of data contains in these different variables.

```
str(salary,give.attr=F)
```

```
## tibble [12,646 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Name                 : chr [1:12646] "AACHOUI, YOUSSEF" "AARNIO, REA T" "ABAJA
S, YASMINA L" "ABARBANELL, JEFFERY S" ...
##  $ campus2              : chr [1:12646] "UNC-CHAPEL HILL" "UNC-CHAPEL HILL" "UNC-C
HAPEL HILL" "UNC-CHAPEL HILL" ...
##  $ dept                 : chr [1:12646] "Microbiology and Immunology" "SW-Research
Projects" "Peds-Hematology/Oncology" "Kenan-Flagler Bus Sch" ...
##  $ position             : chr [1:12646] "Research Professional, Medical" "Function
al Paraprofessional" "Assistant Professor" "Associate Professor" ...
##  $ PRIMARY_WORKING_TITLE: chr [1:12646] "Research Associate" "Graphic Designer" "N
ODESCR" "Associate Professor" ...
##  $ hiredate             : chr [1:12646] "10/10/2011" "1/14/2013" "7/1/2015" "1/1/1
999" ...
##  $ exempt               : chr [1:12646] "Exempt from Personnel Act" "Subject to St
ate Personnel Act" "Exempt from Personnel Act" "Exempt from Personnel Act" ...
##  $ fte                  : num [1:12646] 1 0.8 1 1 1 1 1 1 1 1 ...
##  $ employed             : num [1:12646] 12 12 12 9 12 12 12 9 12 9 ...
##  $ statesal             : logi [1:12646] NA NA NA NA NA NA ...
##  $ nonstsal             : logi [1:12646] NA NA NA NA NA NA ...
##  $ totalsal             : num [1:12646] 49128 33257 139405 181000 41098 ...
##  $ stservyr             : num [1:12646] 1 5 2 20 6 8 6 1 19 1 ...
```

You will notice that the variable “hiredate” is recorded as a character. The following code will first modify the original dataset to change this to a date variable with the format *mm/dd/yyyy*. Then, we will remove the hyphens to create a numeric variable as *yyyymmdd*. Finally, in the spirit of tidyverse, we will convert this data frame to a tibble.

```
salary$hiredate=as.Date(salary$hiredate, format="%m/%d/%Y")
salary$hiredate=as.numeric(gsub("-","",salary$hiredate))
salary=as.tibble(salary)
```

```
## Warning: `as.tibble()` is deprecated as of tibble 2.0.0.
## Please use `as_tibble()` instead.
## The signature and semantics have changed, see `?as_tibble`.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

Now, we will use `head()` to view of first five rows and the modifications made to the original data. The rest of

the assignment will extend off this modified dataset named `salary` which by now should be in your global environment.

```
head(salary,5)
```

```
## # A tibble: 5 x 13
##   Name  campus2 dept  position PRIMARY_WORKING~ hiredate exempt   fte employed
##   <chr> <chr>   <chr> <chr>    <chr>               <dbl> <chr>  <dbl>    <dbl>
## 1 AACH~ UNC-CH~ Micr~ Researc~ Research Associ~ 20111010 Exemp~   1         12
## 2 AARN~ UNC-CH~ SW-R~ Functio~ Graphic Designer 20130114 Subje~   0.8       12
## 3 ABAJ~ UNC-CH~ Peds~ Assista~ NODESCR          20150701 Exemp~   1         12
## 4 ABAR~ UNC-CH~ Kena~ Associa~ Associate Profe~ 19990101 Exemp~   1          9
## 5 ABAR~ UNC-CH~ Inst~ Researc~ Research Techni~ 20110912 Subje~   1         12
## # ... with 4 more variables: statesal <lgl>, nonstsal <lgl>, totalsal <dbl>,
## #   stservyr <dbl>
```

# Assignment

# Part 1: Reducing the Data to a Smaller Set of Interest

## Q1 *(2 Points)*

Create a new dataset named `salary2` that only contains the following variables:

- “Name”
- “dept”
- “position”
- “hiredate”
- “exempt”
- “totalsal”

Then, use the `names()` function to display the variable names of `salary2` .

```
Name = salary$Name
dept = salary$dept
position = salary$position
hiredate = salary$hiredate
exempt = salary$exempt
totalsal = salary$totalsal

salary2 = tibble(Name, dept, position, hiredate, exempt, totalsal)
names(salary2)
```

```
## [1] "Name"     "dept"     "position" "hiredate" "exempt"   "totalsal"
```

# Q2 *(2 Points)*

Now, we modify `salary2`. Rename the variables “dept”,“position”,“exempt”,“totalsal” to “Department”,“Job”,“Exempt”, and “Salary”, respectively. Do this for a new dataset called `salary3` and use `names()` to display the variable names of `salary3`.

```
salary3 = rename(salary2, Department = dept, Job = position, Exempt = exempt, Salary
= totalsal)
names(salary3)
```

```
## [1] "Name"       "Department" "Job"        "hiredate"   "Exempt"
## [6] "Salary"
```

# Q3 *(2 Points)*

Now, we modify `salary3`. Create a new variable called “HireYear” that only contains the first four digits of the variable “hiredate” in a new dataset named `salary4`. *Hint: Use the concept seen in the conversion of flight times to minutes since midnight.* Use the function `str()` to ensure that your new variable “HireYear” reports the year of the date that the employee was hired.

```
salary4 = mutate(salary3, HireYear = hiredate %/% 10000)
str(salary4)
```

```
## tibble [12,646 x 7] (S3: tbl_df/tbl/data.frame)
##  $ Name      : chr [1:12646] "AACHOUI, YOUSSEF" "AARNIO, REA T" "ABAJAS, YASMINA
L" "ABARBANELL, JEFFERY S" ...
##  $ Department: chr [1:12646] "Microbiology and Immunology" "SW-Research Projects"
"Peds-Hematology/Oncology" "Kenan-Flagler Bus Sch" ...
##  $ Job       : chr [1:12646] "Research Professional, Medical" "Functional Paraprof
essional" "Assistant Professor" "Associate Professor" ...
##  $ hiredate  : num [1:12646] 20111010 20130114 20150701 19990101 20110912 ...
##  $ Exempt    : chr [1:12646] "Exempt from Personnel Act" "Subject to State Personn
el Act" "Exempt from Personnel Act" "Exempt from Personnel Act" ...
##  $ Salary    : num [1:12646] 49128 33257 139405 181000 41098 ...
##  $ HireYear  : num [1:12646] 2011 2013 2015 1999 2011 ...
```

# Q4 *(2 points)*

Now, we modify `salary4`. Create a new variable called “YrsEmployed” which reports the number of full years the employee has worked at UNC. Assume that all employees are hired January 1. Create a new dataset named `salary5` and again use `str()` to display the variables in `salary5`. (Use 2020 to create `YrsEmployed`)

```
salary5 = mutate(salary4, YrsEmployed = 2020 - HireYear)
str(salary5)
```

```
## tibble [12,646 x 8] (S3: tbl_df/tbl/data.frame)
##  $ Name       : chr [1:12646] "AACHOUI, YOUSSEF" "AARNIO, REA T" "ABAJAS, YASMINA
L" "ABARBANELL, JEFFERY S" ...
##  $ Department : chr [1:12646] "Microbiology and Immunology" "SW-Research Projects"
"Peds-Hematology/Oncology" "Kenan-Flagler Bus Sch" ...
##  $ Job        : chr [1:12646] "Research Professional, Medical" "Functional Parapro
fessional" "Assistant Professor" "Associate Professor" ...
##  $ hiredate   : num [1:12646] 20111010 20130114 20150701 19990101 20110912 ...
##  $ Exempt     : chr [1:12646] "Exempt from Personnel Act" "Subject to State Person
nel Act" "Exempt from Personnel Act" "Exempt from Personnel Act" ...
##  $ Salary     : num [1:12646] 49128 33257 139405 181000 41098 ...
##  $ HireYear   : num [1:12646] 2011 2013 2015 1999 2011 ...
##  $ YrsEmployed: num [1:12646] 9 7 5 21 9 11 8 4 15 4 ...
```

# Q5 *(4 points)*

Now, we modify `salary5` to create our final dataset named `salary.final`. Use the pipe `%>%` to make the following changes:

- Drop the variables “hiredate” and “HireYear”.
- Sort the observations first by “Department” and then by “YrsEmployed”.
- Rearrange the variables so that “YrsEmployed” and “Salary” are the first two variables in the dataset, in that order, without removing any of the other variables.

After you have used the `%>%` to make these changes, use the function `head()` to display the first 10 rows of `salary.final`.

```
salary.final = select(salary5, Name, Department, Job, Exempt, Salary, YrsEmployed) %
>% arrange(Department, YrsEmployed)
head(salary.final, 10)
```

```
## # A tibble: 10 x 6
##    Name         Department       Job            Exempt         Salary YrsEmployed
##    <chr>        <chr>            <chr>          <chr>           <dbl>       <dbl>
##  1 DALEY, JOS~ A and S - Busin~ Fiscal Affair~ Subject to St~  39646           3
##  2 WEBSTER, C~ A and S - Busin~ HR Coordinator Subject to St~  48814           3
##  3 WOODSON, K~ A and S - Busin~ HR Coordinator Subject to St~  48814           3
##  4 WORTHEN, T~ A and S - Busin~ HR Coordinator Subject to St~  48814           3
##  5 CHESTER, A~ A and S - Busin~ HR Coordinator Subject to St~  47164           4
##  6 GIBSON, JE~ A and S - Busin~ Fiscal Affair~ Subject to St~  47983           4
##  7 RAUSCHER, ~ A and S - Busin~ Fiscal Affair~ Subject to St~  39646           4
##  8 STRINGFELL~ A and S - Busin~ Fiscal Affair~ Subject to St~  39646           4
##  9 WATSON, ST~ A and S - Busin~ HR Coordinator Subject to St~  48814           5
## 10 YOUSEF, HE~ A and S - Busin~ Fiscal Affair~ Subject to St~  47983           5
```

# Part 2: Answering Questions Based on All Data

## Q6 *(2 Points)*

What is the average salary of employees in the Law Department?

Code *(1 Point)*:

```
law_salary = filter(salary.final, Department == "Law")
summary(law_salary)
```

```
##      Name             Department            Job               Exempt
##  Length:112         Length:112         Length:112         Length:112
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      Salary        YrsEmployed
##  Min.   : 33097   Min.   : 3.00
##  1st Qu.: 58659   1st Qu.: 5.00
##  Median : 92095   Median :11.00
##  Mean   :112567   Mean   :12.96
##  3rd Qu.:156261   3rd Qu.:16.25
##  Max.   :400925   Max.   :42.00
```

Answer *(1 Point)*: (Place Answer Here Using Complete Sentences) The average salary of an individuals in the Law Department is $112,567.

## Q7 *(4 Points)*

How many employees have worked in Family Medicine between 5 and 8 years (inclusive) and are exempt from personnel act?

Code *(2 Points)*:

```
empl_fam_med = filter(salary.final, Department == "Family Medicine")
second_emply_fam_med = filter(empl_fam_med, Exempt == "Exempt from Personnel Act")
last_emply_fam_med = filter(second_emply_fam_med, between(YrsEmployed, 5,8))
count(last_emply_fam_med)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    16
```

Answer *(2 Points)*: (Place Answer Here Using Complete Sentences) Exactly 16 employees who have worked in the Family Medicine Department between five to eight years are exempt from the Personnel Act.

# Q8 *(4 Points)*

What is the mean salary of employees from the Linguistics department who are professors, associate professors, or assistant professors?

Code *(2 Points)*:

```
linguis_profs = filter(salary.final, Department == "Linguistics")
linguis_profs_sal = filter(linguis_profs)
linguis_profs_sal_final =filter(linguis_profs_sal, Job == "Professor" | Job == "Assoc
iate Professor" | Job == "Assistant Professor")
summary(linguis_profs_sal_final)
```

```
##      Name             Department            Job               Exempt
##  Length:6           Length:6           Length:6           Length:6
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      Salary       YrsEmployed
##  Min.   :72550   Min.   :11.00
##  1st Qu.:77345   1st Qu.:16.00
##  Median :77789   Median :16.50
##  Mean   :79935   Mean   :16.17
##  3rd Qu.:80766   3rd Qu.:17.75
##  Max.   :92528   Max.   :19.00
```

Answer *(2 Points)*: (Place Answer Here Using Complete Sentences) The mean salary of the employees in the Linguistics Department who are Professors, Associate Professors, and Assistant Professors is 79,935. ## Part 3: Answering Questions Based on Summarized Data

# Q9 *(4 Points)*

Based off the data in `salary.final` , create a grouped summary based off combinations of “Department” and “YrsEmployed”. Call the new tibble `deptyear_summary` . Your summarized tibble, `deptyear_summary` , should report all of the following statistics with corresponding variable names in the following order.

- “n” = number of employees for each combination
- “mean” = average salary for each combination
- “sd” = standard deviation of salary for each combination.
- “min” = minimum salary for each combination.
- “max” = maximum salary for each combination

In the process, make sure you use `ungroup()` with the pipe `%>%` to release the grouping so future work is no longer group specific. Following the creation of `deptyear_summary` , prove that your code worked by using `head()` to view the first 10 rows.

```
deptyear_summary = salary.final%>%
group_by(Department, YrsEmployed)%>%transmute("NumberofEmployees" = n(), "MeanSal" =
mean(Salary), "SdSal" = sd(Salary), min = min(Salary), max = max(Salary)) %>%ungroup
()
head(deptyear_summary, 10)
```

```
## # A tibble: 10 x 7
##    Department             YrsEmployed NumberofEmploye~ MeanSal SdSal   min   max
##    <chr>                        <dbl>            <int>   <dbl> <dbl> <dbl> <dbl>
##  1 A and S - Business Ce~           3                4  46522  4584  39646 48814
##  2 A and S - Business Ce~           3                4  46522  4584  39646 48814
##  3 A and S - Business Ce~           3                4  46522  4584  39646 48814
##  4 A and S - Business Ce~           3                4  46522  4584  39646 48814
##  5 A and S - Business Ce~           4                4  43610. 4589. 39646 47983
##  6 A and S - Business Ce~           4                4  43610. 4589. 39646 47983
##  7 A and S - Business Ce~           4                4  43610. 4589. 39646 47983
##  8 A and S - Business Ce~           4                4  43610. 4589. 39646 47983
##  9 A and S - Business Ce~           5                2  48398.  588. 47983 48814
## 10 A and S - Business Ce~           5                2  48398.  588. 47983 48814
```

# Q10 *(4 Points)*

Using the summarized data in `deptyear_summary`, use the `dplyr` functions to identify the 3 departments that award the lowest average salary for employees who have been employed for 3 years. The output should only show the 3 departments along with the corresponding years employeed, which should all be 3, and the four summarizing statistics created.

Furthermore, explain why the standard deviations for the 3 departments in your list have salary standard deviations of `NA`. What does this mean and how did it occur?

Code *(2 Points)*:

```
deptyear_summary %>%
  filter(YrsEmployed == 3) %>%
  arrange((MeanSal)) %>%
  head(3)
```

```
## # A tibble: 3 x 7
##   Department             YrsEmployed NumberofEmploye~ MeanSal SdSal   min   max
##   <chr>                        <dbl>            <int>   <dbl> <dbl> <dbl> <dbl>
## 1 Religious Studies                3                1   16852    NA 16852 16852
## 2 Ath Olympic Sport Admi~          3                1   19276    NA 19276 19276
## 3 Jewish Studies                   3                1   19750    NA 19750 19750
```

Answer *(2 Points)*: (Place Answer Here Using Complete Sentences) The “NA” for the standard deviation means that the standard deviation cannot be calculated since there is only 1 employee for each of the department for the amount of 3 years. There cannot be a standard deviation with one number. ### Q11 *(4 points)*

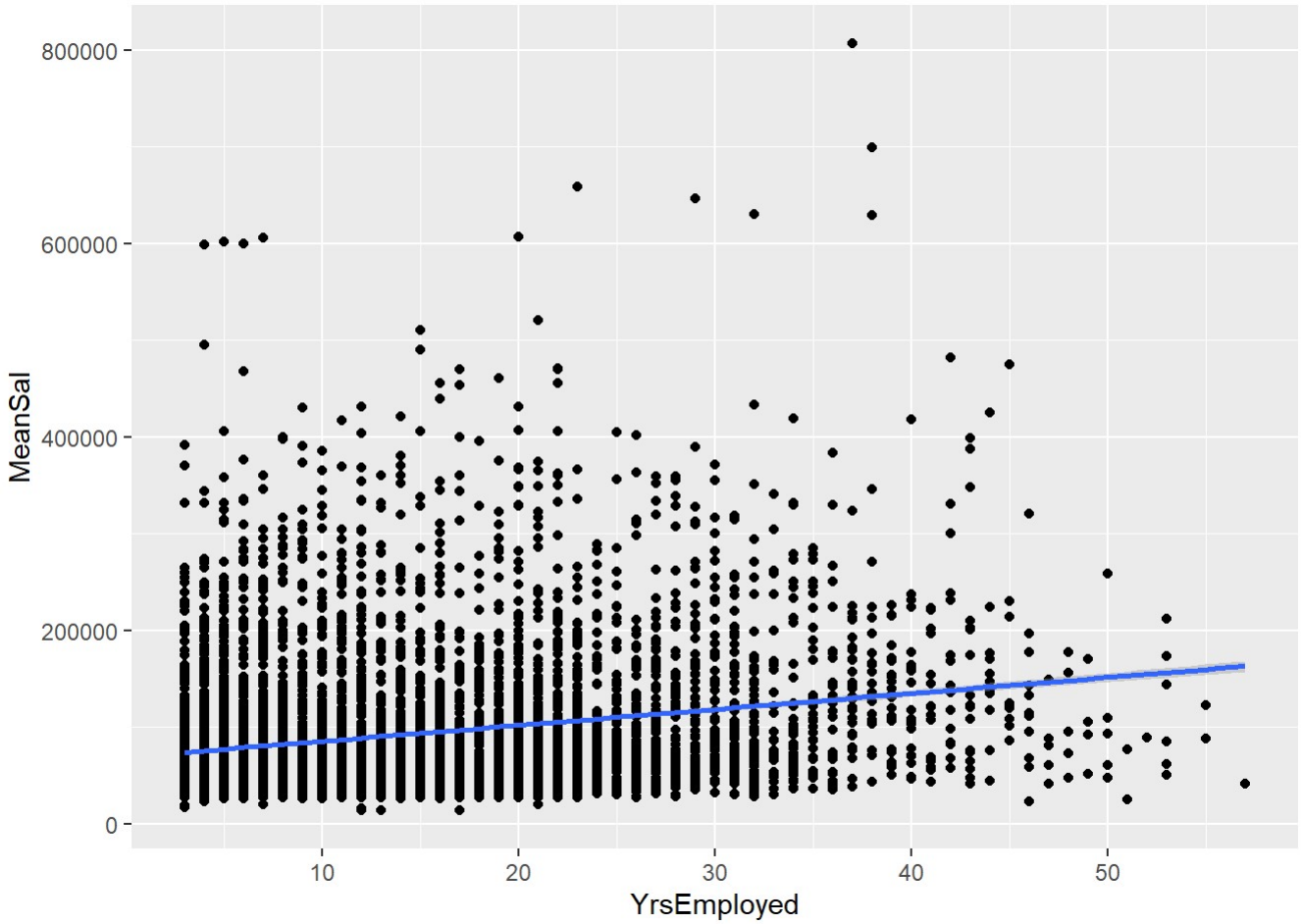
Create a scatter plot using `geom_point()` along with fitted lines using `geom_smooth` with the argument

`method="lm"` showing the linear relationship between average salary and the years employeed. For this plot, use the summarized data in `deptyear_summary`. Following the plot, please explain what this plot suggests about the relationship between the salary a UNC employee makes and how many years that employee has served. Make reference to the figure and use descriptive adjectives (i.e. “strong”, “weak”, etc.) and terms (i.e. “positive”, “negative”, etc.) that are appropriate for discussing linear relationships.

Code and Figure *(2 Points)*:

```
ggplot(data = deptyear_summary) + geom_point(mapping = aes(x = YrsEmployed, y = MeanSal)) + geom_smooth(mapping = aes(x = YrsEmployed, y = MeanSal), method ="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




Answer *(2 Points)*: (Place Answer Here Using Complete Sentences) This plot shows that there is a positive relationship between “YrsEmployed” and “MeanSal”. This means that the average salary of those who have worked at the university for a longer period of time is higher than those who have not worked at the university as long. The longer one works at UNC, the higher their average salary is.

# Q12 *(6 Points)*

The purpose of summarizing the data was to analyze the previously discussed linear relationship by group. In `deptyear_summary`, there are 702 unique departments represented. You can verify this by using `length(unique(deptyear_summary$Department))`. In this part, I want you to select 5 academic

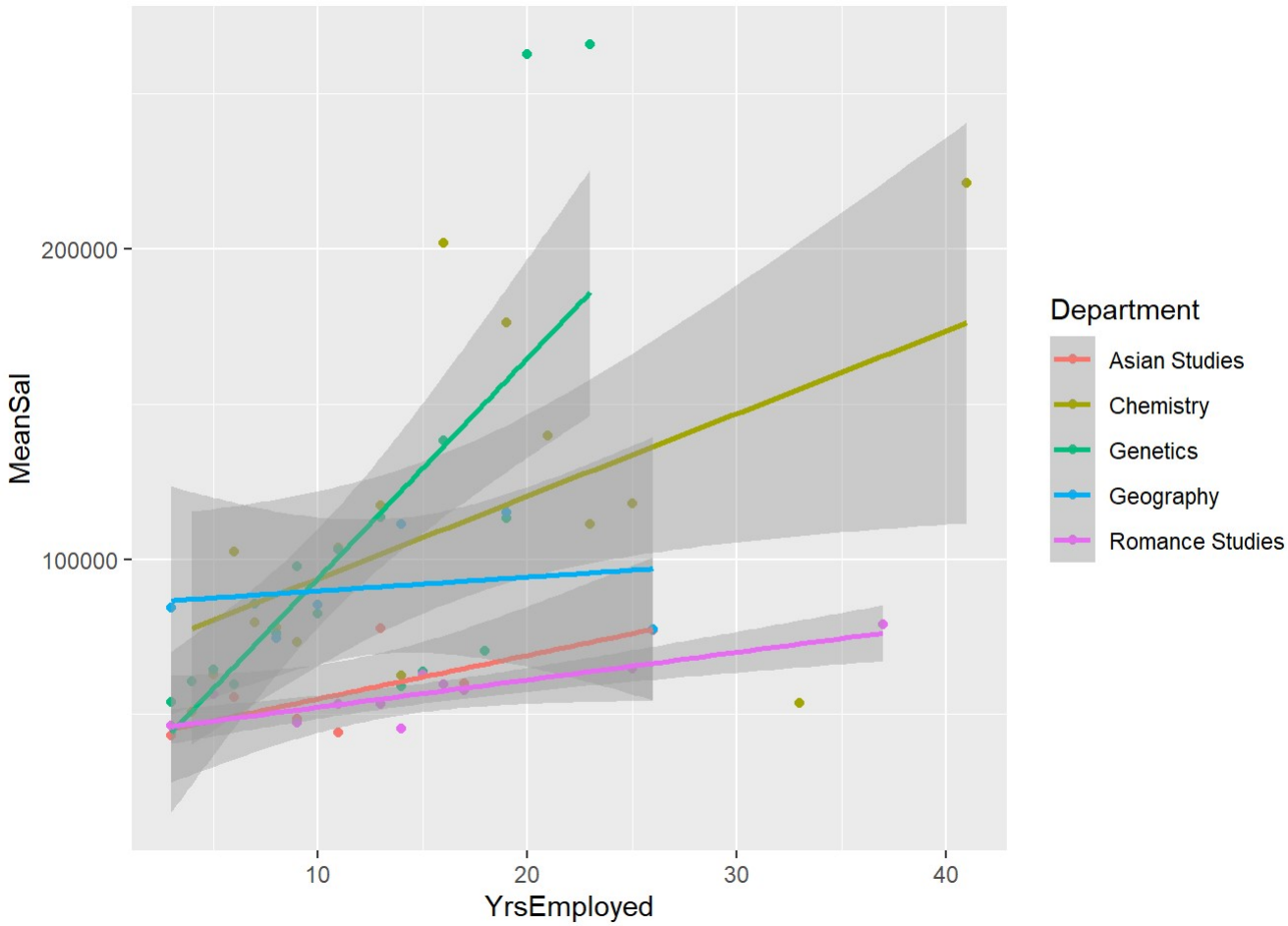
departments, not previously discussed, and in one figure, display the scatter plots and fitted regression lines representing the relationship between average salary and years employed in 5 different colors. Then, in complete sentences, I want you to state what departments you chose and explain the differences and/or similarities between the groups regarding the previously mentioned relationship. Compare departments on the starting salary and the rate of increase in salary based on the fitted lines.

Code and Figure: *(3 Points)*:

```
my_selected_depts = c("Geography","Chemistry","Asian Studies","Romance Studies","Gene
tics")
deptyear_summary%>% filter(Department == my_selected_depts) %>% ggplot(aes( x = YrsEm
ployed, y = MeanSal, color = Department)) + geom_point()+geom_smooth(method = 'lm')
```

```
## Warning in Department == my_selected_depts: longer object length is not a
## multiple of shorter object length
```

```
## `geom_smooth()` using formula 'y ~ x'
```




Answer *(3 Points)*: (Place Answer Here Using Complete Sentences)

The departments I chose were Asian Studies, Chemistry, Genetics, Geography, and Romance Studies. Two of these departments are more so in the Arts, and the other three are in the Sciences. I have noticed that there are not as many people who have been employed long in certain departments, compared to others (whether

that be due to staff leaving/arriving or whenever the departments were created). For example: The Chemistry department has more “Yrs Employed” than the Romance Studies departments. Also, some departments, such as the Geography have a more constant line when it comes to average salary vs years employed. There are also though departments, such as Genetics, that show a less constant and highly variable average salary with more years employed (the more years employed, the mean salary will be much higher).