BUS 316- Final Project

Chaeyon Jang

2023-12-14

Load Packages

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✔ ggplot2 3.4.4     ✔ purrr   1.0.2
## ✔ tibble  3.2.1     ✔ dplyr   1.1.4
## ✔ tidyr   1.3.0     ✔ stringr 1.5.0
## ✔ readr   2.1.2     ✔ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(RMySQL)
```

```
## Loading required package: DBI
```

```
library(knitr)
library(dplyr)
options(scipen=999)
```

Load Data

```
salary_data <- read.csv("https://ballenger.wlu.edu/bus316/State_University_Salary_Data_AY_2015-16.csv")
```

Explore Data

The first step in answering the questions below is to explore the data set using the appropriate exploratory R functions. Use two R functions to explore the data. One of these functions is to list the variables in the data set and provide a sampling of the data. You are to select an appropriate function for the second exploratory function.

```
str(salary_data)
```

```
## 'data.frame':    12287 obs. of  13 variables:
##  $ name    : chr  "AARON, NANCY G" "ABARBANELL, JEFFERY S" "ABARE, BETSY" "ABATE, AAR
ON B" ...
##  $ dept    : chr  "Romance Languages" "Kenan-Flagler Business School" "Institute of M
arine Sciences" "Medicine Administration" ...
##  $ position: chr  "Senior Lecturer" "Associate Professor" "Research Technician" "Acco
unting Technician" ...
##  $ exempt2 : chr  "Exempt" "Exempt" "Subject to State Personnel Act" "Subject to Stat
e Personnel Act" ...
##  $ employed: int  9 9 12 12 12 12 12 12 12 12 ...
##  $ hiredate: int  20030701 19990101 20110912 20090420 20120103 20051003 19960923 2013
0401 19870101 20120702 ...
##  $ fte     : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ status  : chr  "Fixed-Term" "Continuing" "Permanent" "Permanent" ...
##  $ stservyr: int  11 17 3 5 2 15 34 11 27 2 ...
##  $ statesal: int  46350 173000 0 0 41696 56588 41707 0 0 0 ...
##  $ nonstsal: int  0 0 38170 50070 0 4412 0 80227 55803 32889 ...
##  $ totalsal: int  46350 173000 38170 50070 41696 61000 41707 80227 55803 32889 ...
##  $ age     : int  55 57 54 29 35 41 62 36 64 26 ...
```

```
glimpse(salary_data)
```

```
## Rows: 12,287
## Columns: 13
## $ name     <chr> "AARON, NANCY G", "ABARBANELL, JEFFERY S", "ABARE, BETSY", "A…
## $ dept     <chr> "Romance Languages", "Kenan-Flagler Business School", "Instit…
## $ position <chr> "Senior Lecturer", "Associate Professor", "Research Technicia…
## $ exempt2  <chr> "Exempt", "Exempt", "Subject to State Personnel Act", "Subjec…
## $ employed <int> 9, 9, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 9, 9, 12, 12, 9…
## $ hiredate <int> 20030701, 19990101, 20110912, 20090420, 20120103, 20051003, 1…
## $ fte      <dbl> 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1…
## $ status   <chr> "Fixed-Term", "Continuing", "Permanent", "Permanent", "Perman…
## $ stservyr <int> 11, 17, 3, 5, 2, 15, 34, 11, 27, 2, 19, 7, 6, 21, 4, 0, 8, 4,…
## $ statesal <int> 46350, 173000, 0, 0, 41696, 56588, 41707, 0, 0, 0, 107400, 54…
## $ nonstsal <int> 0, 0, 38170, 50070, 0, 4412, 0, 80227, 55803, 32889, 0, 18555…
## $ totalsal <int> 46350, 173000, 38170, 50070, 41696, 61000, 41707, 80227, 5580…
## $ age      <int> 55, 57, 54, 29, 35, 41, 62, 36, 64, 26, 51, 41, 63, 60, 36, 5…
```

Check Data

In a code chuck you are write R tidyverse code to return a distinct list of departments contained in the salary_data data frame and then sort the departments in ascending order. Visually inspect the data looking for potential department misspellings. Use the chuck option, results='hide', to prevent displaying the results of the code chunk when the R markdown document is knitted. In an actual business analytics project, you will do many more quality checks than this.

```
distinct(salary_data, dept) %>%
  arrange(dept)
```

Question 1

Rename the following columns in the salary_data data frame:

stservyr to state_service_yrs statesal to state_salary nonstsal to non_state_salary totalsal to total_salary

Show the structure of the changed data frame.

```
salary_data <- salary_data %>%
  rename(
    "state_service_yrs" = "stservyr",
    "state_salary" = "statesal",
    "non_state_salary" = "nonstsal",
    "total_salary" = "totalsal")
str(salary_data)
```

```
## 'data.frame':    12287 obs. of  13 variables:
##  $ name             : chr  "AARON, NANCY G" "ABARBANELL, JEFFERY S" "ABARE, BETSY" "A
BATE, AARON B" ...
##  $ dept             : chr  "Romance Languages" "Kenan-Flagler Business School" "Insti
tute of Marine Sciences" "Medicine Administration" ...
##  $ position         : chr  "Senior Lecturer" "Associate Professor" "Research Technici
an" "Accounting Technician" ...
##  $ exempt2          : chr  "Exempt" "Exempt" "Subject to State Personnel Act" "Subjec
t to State Personnel Act" ...
##  $ employed         : int  9 9 12 12 12 12 12 12 12 12 ...
##  $ hiredate         : int  20030701 19990101 20110912 20090420 20120103 20051003 1996
0923 20130401 19870101 20120702 ...
##  $ fte              : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ status           : chr  "Fixed-Term" "Continuing" "Permanent" "Permanent" ...
##  $ state_service_yrs: int  11 17 3 5 2 15 34 11 27 2 ...
##  $ state_salary     : int  46350 173000 0 0 41696 56588 41707 0 0 0 ...
##  $ non_state_salary : int  0 0 38170 50070 0 4412 0 80227 55803 32889 ...
##  $ total_salary     : int  46350 173000 38170 50070 41696 61000 41707 80227 55803 328
89 ...
##  $ age              : int  55 57 54 29 35 41 62 36 64 26 ...
```

Question 2

Determine the mean total salary of employees in the Neurosurgery department. Name the variable mean_total_salary. You are to generate two versions of your output:

Return a data frame consisting of a single column containing the variable name and the mean total salary.

```
mean_total_salary <- mean(salary_data$total_salary[salary_data$dept == "Neurosurgery"])

neuro_sal_2 <- data.frame(mean_total_salary = mean_total_salary)
print(neuro_sal_2)
```

```
##   mean_total_salary
## 1          380058.1
```

Return a data frame consisting of two columns. The first is to be the name of the department and the second the mean total salary:

```
mean_total_salary <- mean(salary_data$total_salary[salary_data$dept == "Neurosurgery"])

neuro_sal <- data.frame(variable_name = "mean_total_salary", mean_total_salary = mean_to
tal_salary)
print(neuro_sal)
```

```
##       variable_name mean_total_salary
## 1 mean_total_salary          380058.1
```

Question 3

Create a data frame named fulltime that includes only full-time employees and the following columns from the salary-data data set: name, dept, position, age, status, state_salary, non_state_salary, and total_salary. Output the first 15 records of the fulltime data frame.

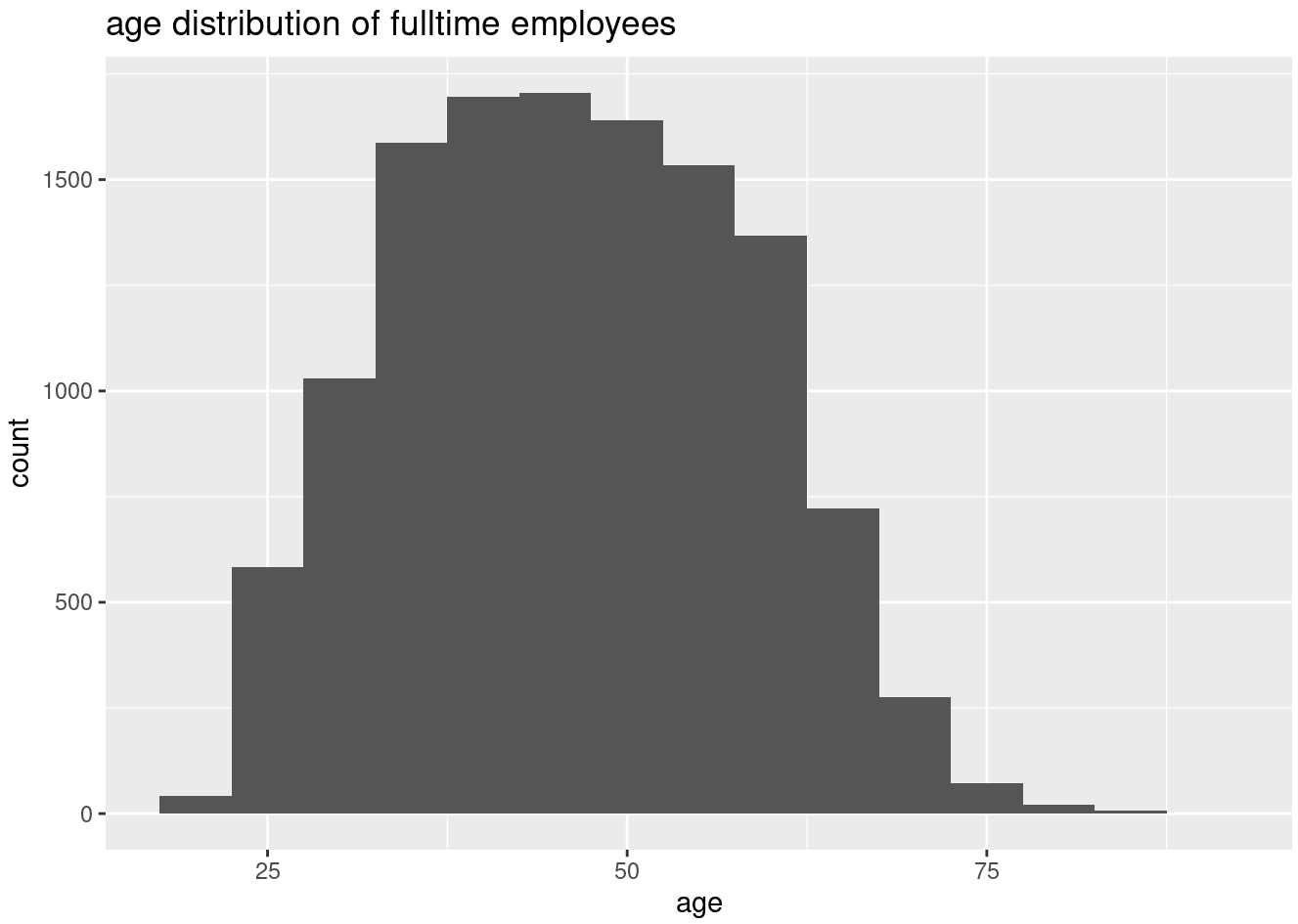
```
fulltime <- salary_data %>%
  select(name, dept, position, age, status, state_salary, non_state_salary, total_salar
y) %>%
  filter("fte" >= 1)
```

```
head(fulltime, 15)
```

```
##                         name                          dept
## 1             AARON, NANCY G             Romance Languages
## 2      ABARBANELL, JEFFERY S Kenan-Flagler Business School
## 3               ABARE, BETSY  Institute of Marine Sciences
## 4             ABATE, AARON B       Medicine Administration
## 5         ABATEMARCO, JODI M           School of Education
## 6  ABBOTT-LUNSFORD, SHELBY L       Medicine Administration
## 7         ABBOTTS, WILLIAM C                       Biology
## 8          ABDOULAYI, SARA M    Carolina Population Center
## 9            ABDULLAH, LUBNA    Cys Fibrosis/Pulmonary Res
## 10                ABE, PAIGE         Housing Res Education
## 11         ABELS, KIMBERLY T               Writing Center
## 12           ABERG, CERESA M              Human Resources
## 13     ABERNATHY, PENELOPE M Journalism/Mass Communication
## 14             ABLE, HARRIET           School of Education
## 15    ABOYADE-COLE, AYOOLA A   Comprehensive Cancer Center
##                                    position age     status state_salary
## 1                           Senior Lecturer  55 Fixed-Term        46350
## 2                       Associate Professor  57 Continuing       173000
## 3                       Research Technician  54  Permanent            0
## 4                     Accounting Technician  29  Permanent            0
## 5                 Student Services Assistant  35  Permanent        41696
## 6                               HR Consultant  41  Permanent        56588
## 7                     Accounting Technician  62  Permanent        41707
## 8         Research Associate-Project Manager  36 Continuing            0
## 9                         Research Associate  64 Continuing            0
## 10                        Community Director  26 Continuing            0
## 11 DIRECTOR, WRITING CENTER and LEARNING CE  51 Continuing       107400
## 12            Staffing Support Services Mgr  41  Permanent        54445
## 13                               Professor  63 Continuing       101706
## 14                            Assoc. Prof.  60 Continuing        89356
## 15                     Regulatory Associate  36  Permanent            0
##    non_state_salary total_salary
## 1                 0        46350
## 2                 0       173000
## 3             38170        38170
## 4             50070        50070
## 5                 0        41696
## 6              4412        61000
## 7                 0        41707
## 8             80227        80227
## 9             55803        55803
## 10            32889        32889
## 11                0       107400
## 12            18555        73000
## 13            50094       151800
## 14             9929        99285
## 15            57854        57854
```

Discuss the distribution of the age variable for full-time employees.

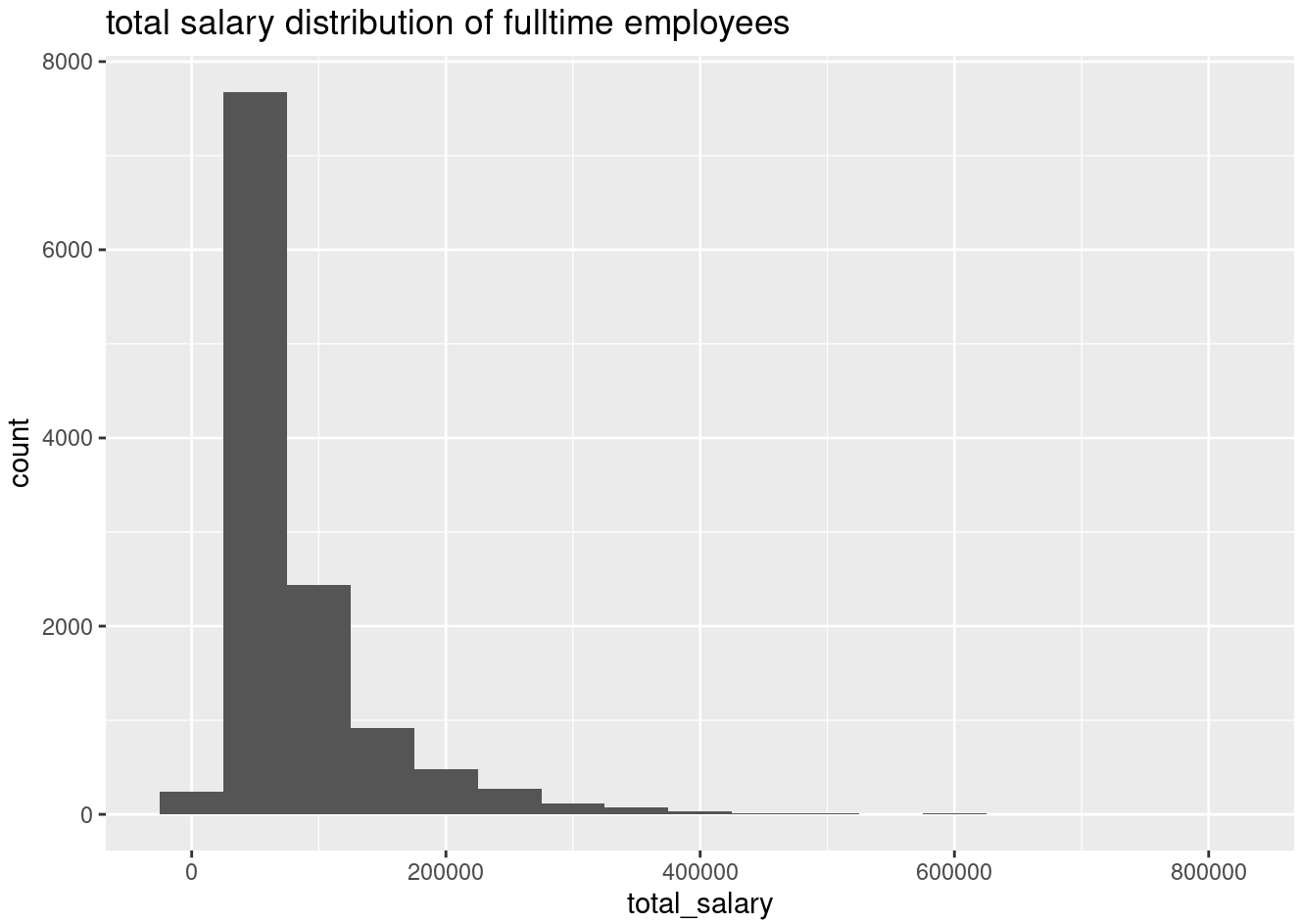
```
ggplot(data = fulltime, mapping = aes(x = age)) +
  geom_histogram(binwidth = 5) +
  labs(title = "age distribution of fulltime employees")
```



The distribution of full time employees is rather spread out; there is not a distinct shape that can really describe the distribution, where the graph could look multimodal, with multiple peaks throughout the graph. The data shows that the age range is from 20 to past 80, with majority in the 30-60 age range.

Discuss the distribution of the total_salary variable for full-time employees.

```
ggplot(data = fulltime, mapping = aes(x = total_salary)) +
  geom_histogram(binwidth = 50000) +
  labs(title = "total salary distribution of fulltime employees")
```

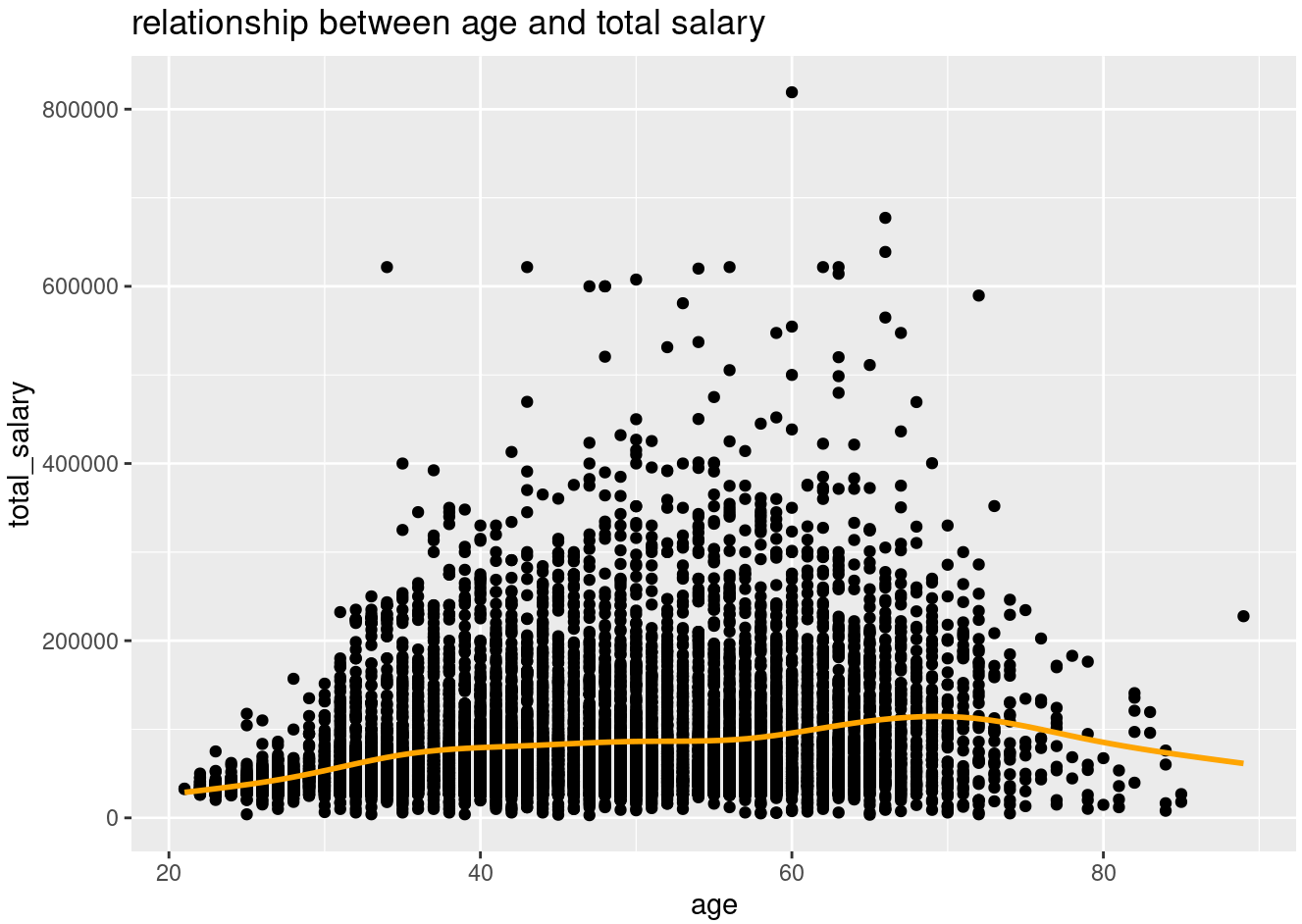


The graph is very right skewed, showing that most full time employees make around 0-$150,000.

Discuss the relationship of age to total_salary for full-time employees. Add a smooth trend line to your plot. Suppress outputting warnings and messages.

```
suppressWarnings({
  ggplot(fulltime, mapping = aes(x = age, y = total_salary)) +
    scale_color_continuous(low = "red", high = "blue") +
    geom_point() +
    geom_smooth(color = "orange", se = FALSE) +
    labs(title = "relationship between age and total salary")
})
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



The scatter plot shows that there is a slight correlation of increasing age and higher total salary, which can be seen also looking at the trend line. This could be explained by the possibility that most people who take higher paying positions tend to be older because they relatively have more experience and connections.

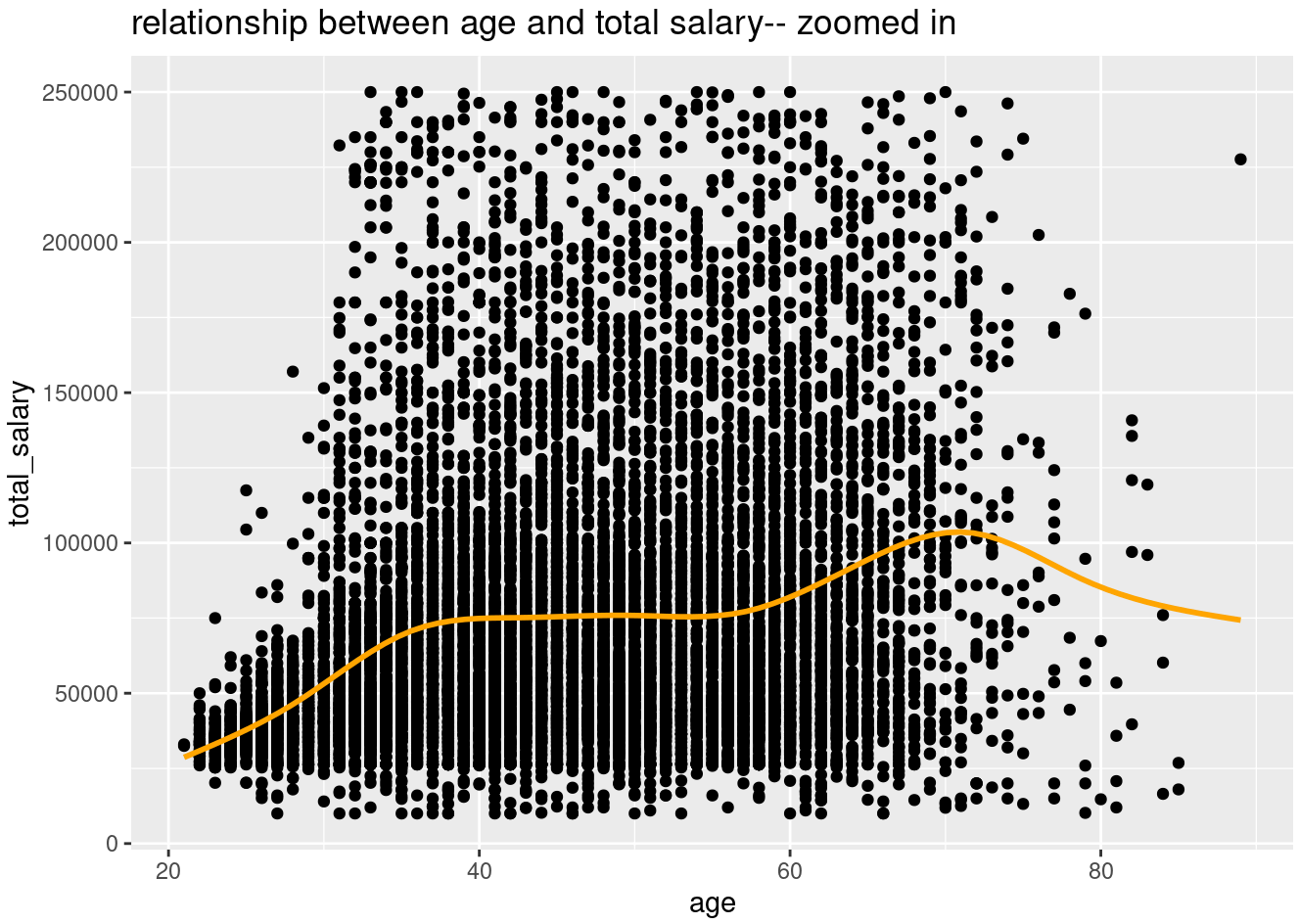
Repeat the previous plot by zooming in on the y axis and output only total salaries between $10,000 and $250,000. Add a smooth trend line to your plot. Suppress outputting warnings and messages.

```
suppressWarnings({
ggplot(fulltime, mapping = aes(x = age, y = total_salary)) +
  scale_color_continuous(low = "red", high = "blue") +
  geom_point() +
  geom_smooth(color = "orange", se = FALSE) +
  ylim(10000, 250000) +
  labs(title = "relationship between age and total salary-- zoomed in")
})
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 404 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 404 rows containing missing values (`geom_point()`).
```



Question 4

Using the fulltime data frame, return a data frame of Neurosurgery department employees earning more than $500,000.

```
n_pay <- fulltime %>%
  select(name, dept, position, age, status, state_salary, non_state_salary, total_salar
y) %>%
  filter(dept == "Neurosurgery", total_salary > 500000)
print(n_pay)
```

```
##                    name         dept                      position age
## 1    CAMPBELL, DENNIS M Neurosurgery  Adjunct Assistant Professor  34
## 2       CARSON, LARRY V Neurosurgery           Clinical Professor  63
## 3      EWEND, MATTHEW G Neurosurgery                      DIRECTOR  50
## 4     JAUFMANN, BRUCE P Neurosurgery Clinical Associate Professor  56
## 5 KILPATRICK, MICHAUX R Neurosurgery Clinical Assistant Professor  43
## 6        WADON, CAROL M Neurosurgery Clinical Associate Professor  56
##       status state_salary non_state_salary total_salary
## 1 Fixed-Term            0           621722       621722
## 2 Fixed-Term            0           621722       621722
## 3 Continuing            0           607648       607648
## 4 Fixed-Term            0           621722       621722
## 5 Fixed-Term            0           621722       621722
## 6 Fixed-Term            0           505390       505390
```

Why are these professors paid so well? These professors are paid well because they have reached a high position in a very difficult department; you can see that there are only 6 people who have managed to specialize in neurosurgery, making the occupation more valuable.

Question 5

Subset the fulltime data frame, by creating a new data frame that contains only the full time employees of the Radiology department. Order the employees by total salary with the highest paid employee appearing first. Name the new data frame, radiology_dept. Once created display the first 10 observations.

```
radiology_dept <- fulltime %>%
  select(name, dept, position, age, status, state_salary, non_state_salary, total_salar
y) %>%
  filter(dept == "Radiology") %>%
  arrange(desc(total_salary))
print(radiology_dept)
```

```
##                      name      dept                        position age
## 1         MAURO, MATTHEW A Radiology                        DIRECTOR  63
## 2            LEE, JOSEPH K Radiology                       Professor  67
## 3         BURKE, CHARLES T Radiology     Clinical Associate Professor  44
## 4           MOLINA, PAUL L Radiology                       Professor  56
## 5         STAVAS, JOSEPH M Radiology              Clinical Professor  59
## 6          DIXON, ROBERT G Radiology     Clinical Associate Professor  55
## 7       CASTILLO, MAURICIO Radiology                       Professor  55
## 8       SEMELKA, RICHARD C Radiology                       Professor  54
## 9               SMITH, J K Radiology            Professor with Tenure  52
## 10       FIELDING, JULIA R Radiology              Associate Professor  53
## 11        RENNER, JORDAN B Radiology                       Professor  59
## 12         WONG, TERENCE Z Radiology                       Professor  59
## 13        SOLANDER, STEN Y Radiology     Clinical Associate Professor  55
## 14      WARSHAUER, DAVID M Radiology                       Professor  61
## 15              LIN, WEILI Radiology    Dixie Lee Boney Soo Professor  50
## 16         FORDHAM, LYNN A Radiology              Associate Professor  51
## 17          CLARKE, JOHN P Radiology     Clinical Associate Professor  63
## 18      JEWELLS, VALERIE L Radiology     Clinical Associate Professor  53
## 19            CHONG, WUI K Radiology     Clinical Associate Professor  57
## 20       HYSLOP, WILLIAM B Radiology     Clinical Associate Professor  54
## 21       KUZMIAK, CHERIE M Radiology              Associate Professor  46
## 22        KOOMEN, MARCIA A Radiology     Clinical Associate Professor  66
## 23   BIRCHARD, KATHERINE R Radiology     Clinical Assistant Professor  40
## 24       HUANG, BENJAMIN Y Radiology              Assistant Professor  41
## 25              KIM, KYUNG Radiology     Clinical Assistant Professor  43
## 26               YU, HYEON Radiology     Clinical Assistant Professor  47
## 27         LURY, KENNETH M Radiology           Clinical Assist. Prof.  60
## 28       NISSMAN, DANIEL B Radiology     Clinical Assistant Professor  46
## 29       PARKER, LEONARD A Radiology              Associate Professor  70
## 30          SHEN, DINGGANG Radiology            Professor with Tenure  45
## 31         ISAACSON, ARI J Radiology     Clinical Assistant Professor  36
## 32        JORDAN, SHERYL G Radiology     Clinical Associate Professor  55
## 33        KHANDANI, AMIR H Radiology  Associate Professor with Tenure  50
## 34           LEE, SHEILA S Radiology     Clinical Assistant Professor  38
## 35         BURKE, LAUREN M Radiology     Clinical Assistant Professor  33
## 36          HARTMAN, HEIDI Radiology     Clinical Assistant Professor  33
## 37             LEE, YUEH Z Radiology              Assistant Professor  41
## 38            MEHTA, NISHA Radiology     Clinical Assistant Professor  32
## 39     NORTHAM, MEREDITH C Radiology     Clinical Assistant Professor  33
## 40       SAMS, CASSANDRA M Radiology     Clinical Assistant Professor  33
## 41              HAN, TAE IL Radiology     Clinical Assistant Professor  48
## 42     COLLICHIO, ROBERT J Radiology  Assoc Chair for Admin/Radiology  61
## 43       HEYNEMAN, LAURA E Radiology     Clinical Associate Professor  51
## 44                LI, ZIBO Radiology              Associate Professor  36
## 45        IVANOVIC, MARIJA Radiology     Clinical Associate Professor  62
## 46              SMITH, H. E Radiology     Research Associate Professor  40
## 47    MCCARTNEY, WILLIAM H Radiology                       Professor  69
## 48         WILCOX, CLAIRE B Radiology     Clinical Associate Professor  68
## 49             LEE, ELLIE R Radiology     Clinical Assistant Professor  44
## 50       CRAWFORD, THOMAS J Radiology               Systems Specialist  55
## 51     HENDERSON, LOUISE M Radiology              Assistant Professor  40
```

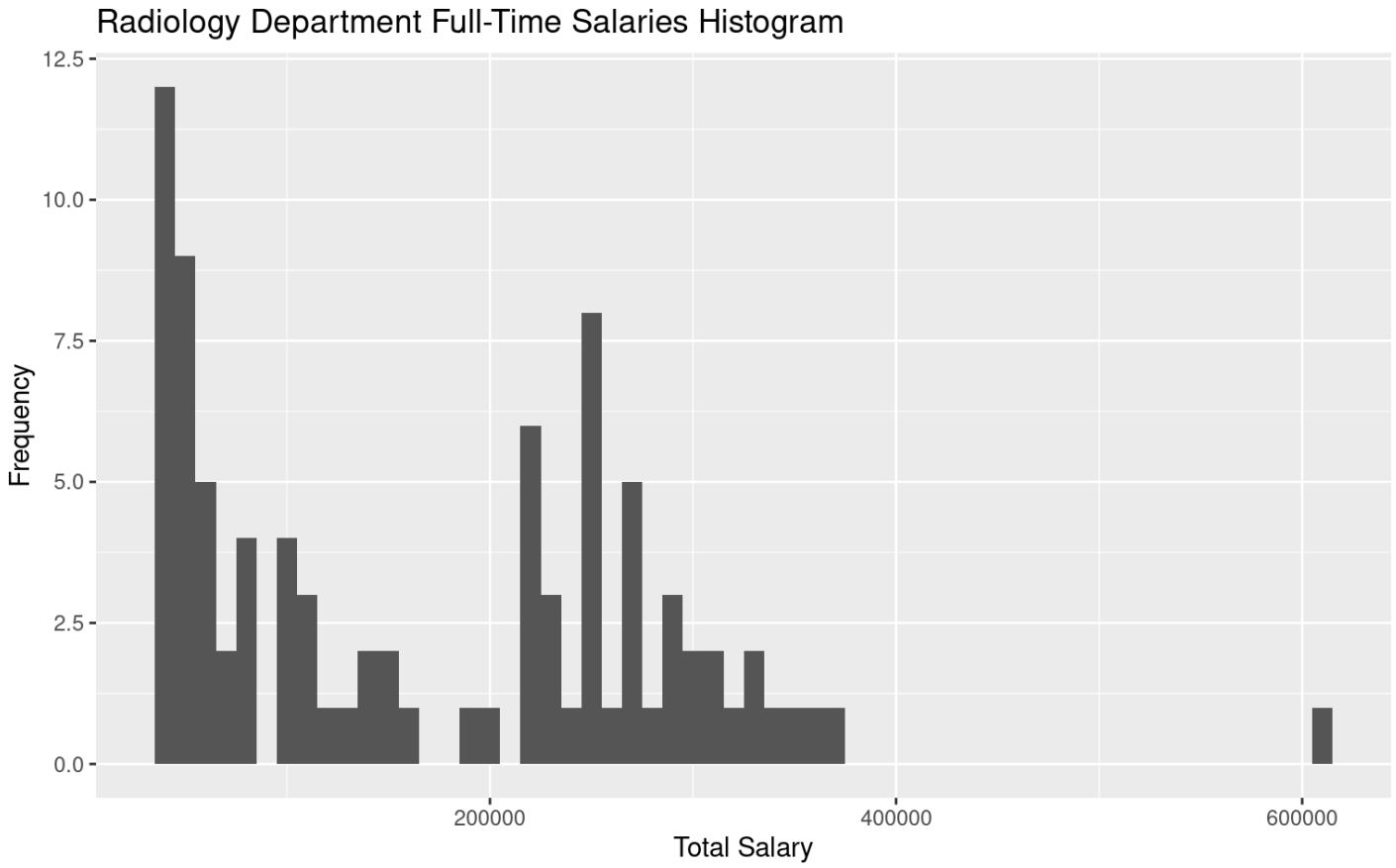
```
## 52      ALVAREZ, HORTENSIA Radiology              Clinical Professor  57
## 53      BOUGHTON, DANIEL J Radiology                Business Officer  41
## 54              AN, HONGYU Radiology             Assistant Professor  45
## 55      PARROTT, MATTHEW C Radiology             Assistant Professor  37
## 56            SHEIKH, ARIF Radiology     Clinical Assistant Professor  46
## 57              YUAN, HONG Radiology     Research Assistant Professor  41
## 58                GAO, WEI Radiology             Assistant Professor  32
## 59            WU, ZHANHONG Radiology     Research Assistant Professor  39
## 60          YAP, PEW THIAN Radiology             Assistant Professor  36
## 61       BENEFIELD, THAD S Radiology                     Statistician  38
## 62       PETRIN, FERNAND H Radiology         Business Systems Analyst  57
## 63       HOLLAND, VICKIE E Radiology                     HR Associate  49
## 64     CREIGHTON, ANGELA H Radiology         Contracts/Grants Manager  43
## 65           AKER, DIXIE K Radiology                  Systems Analyst  45
## 66          USSERY, LISA A Radiology              Accounting Manager  50
## 67          KIRK, SHANAH R Radiology             Research Specialist  50
## 68 RAMALHO, JORGE MIGUEL P Radiology             Research Instructor  40
## 69               SHI, FENG Radiology  Postdoctoral Research Associate  34
## 70             WU, GUORONG Radiology               POST-DOC RES ASSOC  36
## 71      BOWEN, ELIZABETH A Radiology             Executive Assistant  44
## 72           MARSH, MARY W Radiology              Research Associate  27
## 73           STEED, DOREEN Radiology           Research Mammographer  50
## 74           NESBITT, ANNE Radiology         Admin Support Specialist  50
## 75         PRICE, CHERIE L Radiology                     HR Associate  57
## 76        CLARK, MICHELE L Radiology         Admin Support Specialist  45
## 77         KNOP, GABRIEL F Radiology   Social/Clinical Research Spec.  30
## 78          ARMAO, DIANE M Radiology             Research Instructor  59
## 79     BOOMHOWER, JEREMY D Radiology         Admin. Support Associate  38
## 80      CARVER, VIRGINIA B Radiology         Admin. Support Associate  39
## 81         HAUSER, JASON M Radiology         Admin. Support Associate  41
## 82        HARTMAN, TERRY S Radiology   Social/Clinical Research Asst.  26
## 83       MELVILLE, WILMA C Radiology      Administrative Secretary II  58
## 84          BARBAL, ISABEL Radiology         Admin. Support Associate  57
## 85      PENDER, JENNIFER L Radiology            Accounting Technician  39
## 86      BIRDSONG, LAURIE B Radiology Public Communications Specialist  40
## 87     FISCHER, MICHELLE C Radiology         Admin. Support Associate  25
## 88        HOOTS, TIFFANY N Radiology   Social/Clinical Research Asst.  31
##                       status state_salary non_state_salary total_salary
## 1                 Continuing            0           614176       614176
## 2                 Continuing            0           375000       375000
## 3                 Fixed-Term            0           365000       365000
## 4                 Continuing        15745           334255       350000
## 5                 Fixed-Term            0           345000       345000
## 6                 Fixed-Term            0           335000       335000
## 7                 Continuing        15745           316255       332000
## 8                 Continuing        15745           306255       322000
## 9                 Continuing        17813           292187       310000
## 10                Continuing        15745           294005       309750
## 11                Continuing            0           300000       300000
## 12                Fixed-Term        75745           224255       300000
## 13                Fixed-Term        15745           279255       295000
## 14                Continuing        23005           271995       295000
```

```
## 15              Continuing        22735          263805       286540
## 16              Continuing        15745          269255       285000
## 17              Fixed-Term            0          275000       275000
## 18              Fixed-Term        15745          259255       275000
## 19              Fixed-Term        15745          254255       270000
## 20              Fixed-Term        15745          254255       270000
## 21              Continuing        15745          254255       270000
## 22              Fixed-Term        15745          244255       260000
## 23              Fixed-Term        20745          234255       255000
## 24              Fixed-Term        15745          239255       255000
## 25              Fixed-Term            0          251304       251304
## 26              Fixed-Term            0          251000       251000
## 27              Fixed-Term        15745          234255       250000
## 28              Fixed-Term        50446          199554       250000
## 29              Continuing        15745          234255       250000
## 30              Continuing        75589          174411       250000
## 31              Fixed-Term            0          240000       240000
## 32              Fixed-Term        15745          219255       235000
## 33              Continuing        15745          214255       230000
## 34              Fixed-Term            0          230000       230000
## 35              Fixed-Term        15745          204255       220000
## 36              Fixed-Term        15745          204255       220000
## 37              Fixed-Term        85622          134378       220000
## 38              Fixed-Term            0          220000       220000
## 39              Fixed-Term            0          220000       220000
## 40              Fixed-Term            0          220000       220000
## 41              Fixed-Term        15745          180969       196714
## 42              Continuing        10000          181064       191064
## 43              Fixed-Term        11809          149191       161000
## 44              Fixed-Term            0          150000       150000
## 45              Fixed-Term            0          149968       149968
## 46              Fixed-Term        72500           72500       145000
## 47              Fixed-Term         7873          134627       142500
## 48              Fixed-Term       131250               0       131250
## 49              Fixed-Term        11809          108191       120000
## 50               Permanent            0          114698       114698
## 51              Fixed-Term            0          109900       109900
## 52              Fixed-Term            0          109000       109000
## 53               Permanent         4885           96717       101602
## 54              Fixed-Term            0          100000       100000
## 55              Fixed-Term         4945           93965        98910
## 56              Fixed-Term        15745           81755        97500
## 57              Fixed-Term            0           85000        85000
## 58              Fixed-Term        51000           29000        80000
## 59              Fixed-Term        80000               0        80000
## 60              Fixed-Term        53333           26667        80000
## 61              Continuing            0           70840        70840
## 62               Permanent            0           68336        68336
## 63               Permanent        17484           44866        62350
## 64               Permanent            0           61719        61719
## 65               Permanent            0           60000        60000
## 66               Permanent            0           58602        58602
```

```
## 67                  Permanent            0            55940        55940
## 68 Temporary/Visiting Faculty        55000                0        55000
## 69                 Fixed-Term            0            55000        55000
## 70                 Fixed-Term            0            55000        55000
## 71                  Permanent            0            53987        53987
## 72                 Continuing            0            50000        50000
## 73                  Permanent            0            48564        48564
## 74                  Permanent            0            48446        48446
## 75                  Permanent            0            48446        48446
## 76                  Permanent            0            48349        48349
## 77                  Permanent            0            45000        45000
## 78                 Fixed-Term            0            43546        43546
## 79                  Permanent            0            42593        42593
## 80                  Permanent            0            42593        42593
## 81                  Permanent            0            42593        42593
## 82                  Permanent            0            42168        42168
## 83                  Permanent            0            41789        41789
## 84                  Permanent            0            40061        40061
## 85                  Permanent            0            37690        37690
## 86                  Permanent            0            37681        37681
## 87                  Permanent            0            37142        37142
## 88                  Permanent            0            36360        36360
```

Next, in a separate code chuck, create a histogram of Radiology Department full time salaries. The histogram plot is to be centered, 5″ high and 8″ wide.

```
ggplot(data = radiology_dept, mapping = aes(x = total_salary)) +
  geom_histogram(binwidth = 10000) +
  labs(x = "Total Salary", y = "Frequency",
       title = "Radiology Department Full-Time Salaries Histogram")
```



Discuss the distribution of salaries in the Radiology Department.

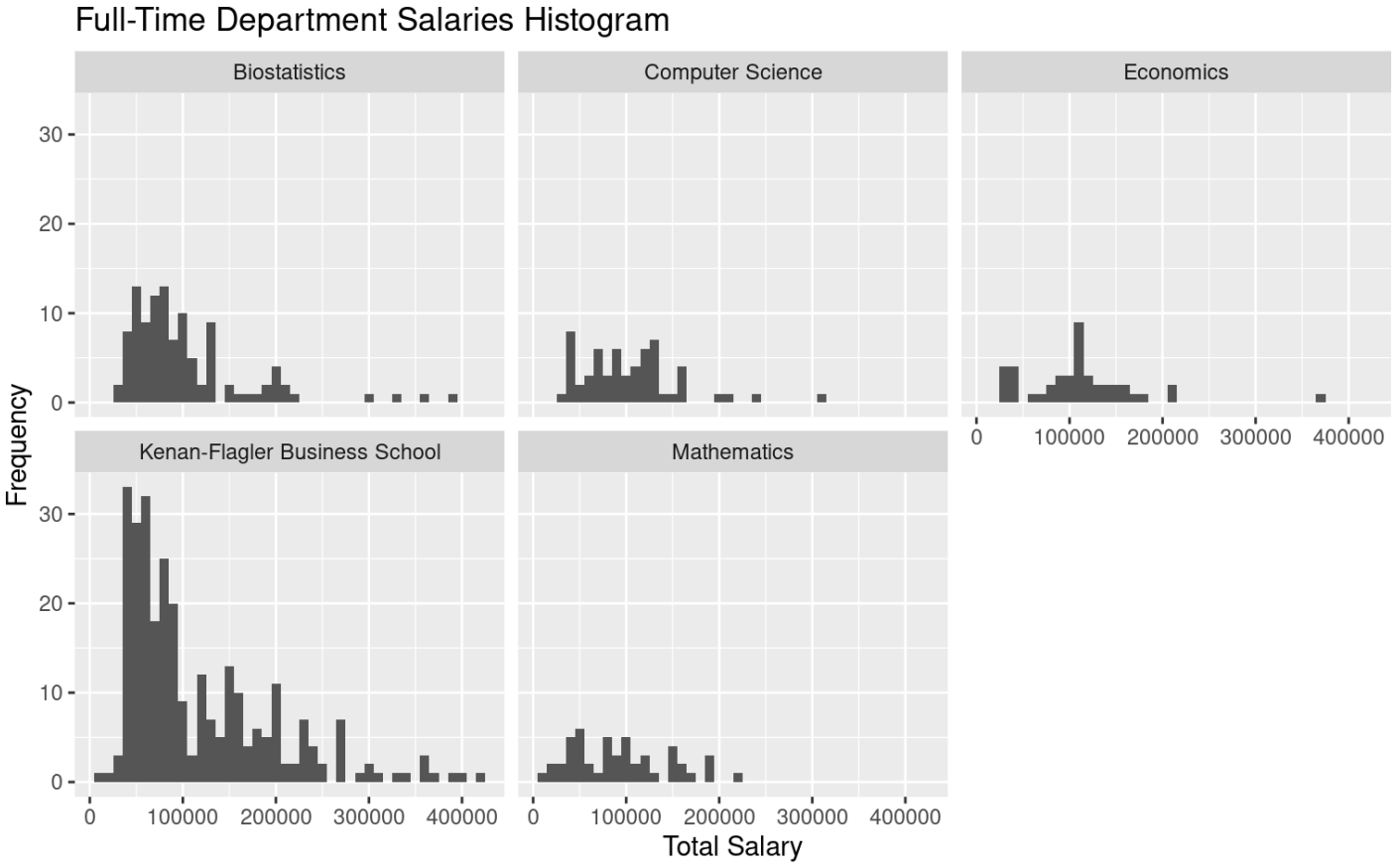
The salary distribution for the radiology department is bimodal, where the concentration of data is in two peaks. Overall, the graph is right-skewed, with majority of the total salary data in the radiology department is on the lower end, from about $30,000 to $175,000 and from $200,000 to $375,000.

What are some some explanations of the shape? Some reasons why there might be this distribution is from high quantities of people in radiology, depsite the occupation being not as competitive and difficult as neurosurgery. The bimodal distribution could be explained by leaps in positions, going from an employee in the department to having a higher-paying position with authority.

Question 6

Use faceting to create histograms of full-time department salaries (use fulltime data frame) for Biostatistics, Computer Science, Economics, Kenan-Flagler Business School, Mathematics, Statistics and Operations Res.

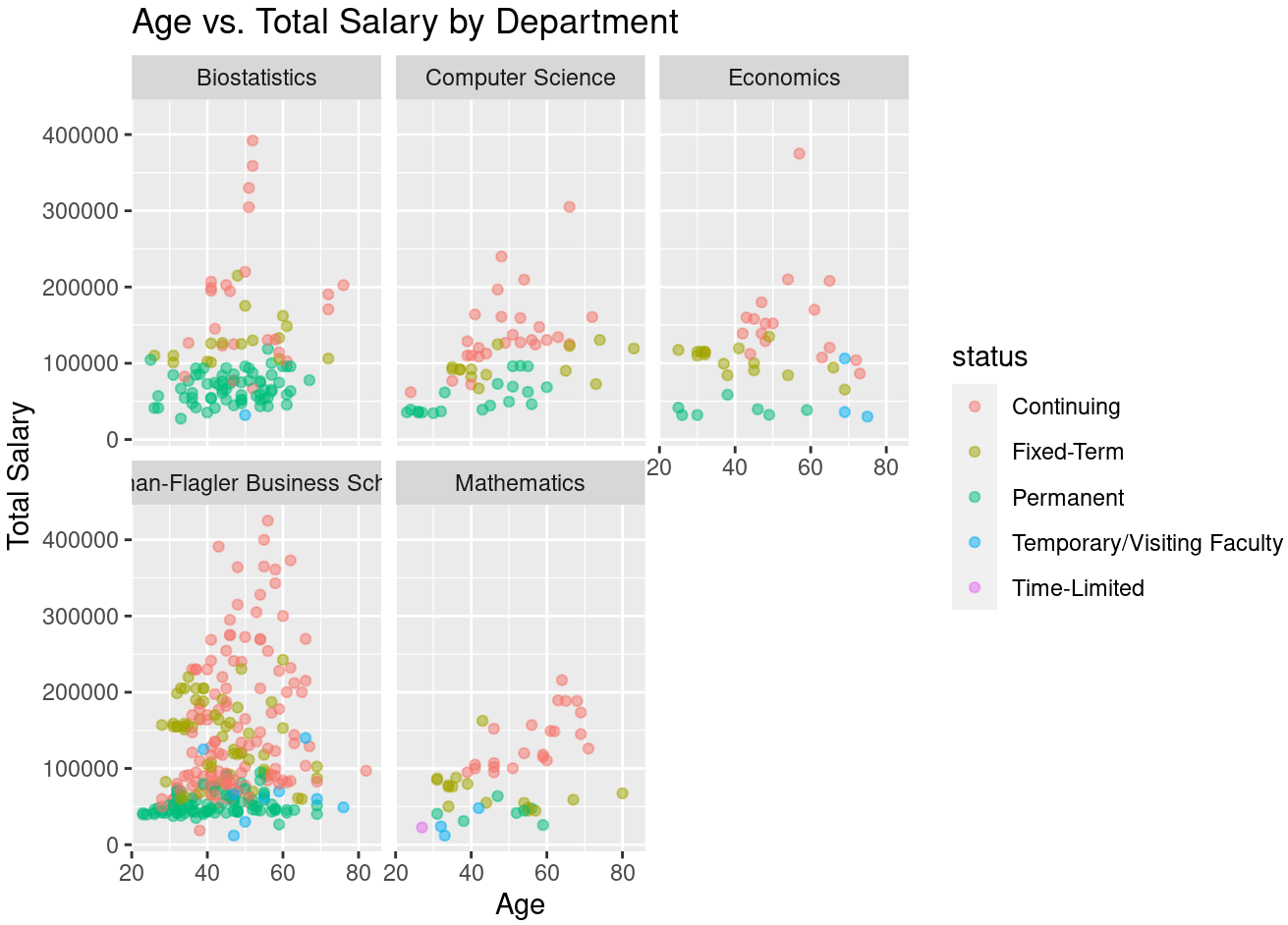
```
selected_dept <- c("Biostatistics", "Computer Science", "Economics",
                   "Kenan-Flagler Business School", "Mathematics",
                   "Statistics", "Operations Res")


filter_dept_data <- fulltime %>%
  filter(dept %in% selected_dept)

ggplot(filter_dept_data, aes(x = total_salary)) +
  geom_histogram(binwidth = 10000) +
  facet_wrap(~ dept) +
  labs(x = "Total Salary", y = "Frequency",
       title = "Full-Time Department Salaries Histogram")
```



Discuss the distribution of salaries in and across the six departments. It seems that all data sets are between 0 - $200,000 average; there is a lot more spread across the business department, which makes sense because the business department is very diverse in terms of specialization–there is also a lot more people in the business school that also are not as specialized, resulting in lower pay. The other departments seem to be have around the same range of average total salary of 0-$200,000– the spread seems more wide across the other departments because there are less employees in these departments.

Use faceting to plot age vs. total salary for the same six departments and color points using the status variable. Finally, adjust the transparency to .5.

```
filter_dept_data <- fulltime %>%
  filter(dept %in% selected_dept)

ggplot(filter_dept_data, aes(x = age, y = total_salary, color = status)) +
  geom_point(alpha = 0.5) +
  facet_wrap(~ dept) +
  labs(x = "Age", y = "Total Salary", title = "Age vs. Total Salary by Department")
```



Question 7

Create a data frame called dept_summary whose rows are the departments and whose columns are: department size, mean department salary, median department salary, and maximum salary (use total_salary for salary).

```
dept_summary <- fulltime %>%
  group_by(dept) %>%
  summarise(
    department_size = n(),
    mean_dept_salary = mean(total_salary),
    median_dept_salary = median(total_salary),
    max_salary = max(total_salary)
  ) %>%
  ungroup()

print(dept_summary)
```

```
## # A tibble: 304 × 5
##    dept           department_size mean_dept_salary median_dept_salary max_salary
##    <chr>                    <int>            <dbl>              <dbl>      <int>
##  1 AHEC Support-…              26           69789.             64533      135193
##  2 Acad Sup Prog…              15           55798.             50600      115000
##  3 Academic Advi…              42           49985.             45000      109625
##  4 Accounting Se…              17           57417.             59342      103306
##  5 Ackland Art M…              19           51543.             41000      140050
##  6 Admissions                  46           57487.             49000      195700
##  7 African Studi…               2           35970              35970       43475
##  8 African, Afri…              23           65170.             68000      135608
##  9 Airport                      1           47351              47351       47351
## 10 Alcohol Studi…              16           49232.             49180.      84685
## # ℹ 294 more rows
```

Create an new data frame, top_dept_means, that includes two columns, dept and mean_dept_total_salary. The new data frame is to order the departments with the highest mean total salary appearing first and restrict the departments to the 10 highest paid.

```
top_dept_means <- dept_summary %>%
  arrange(desc(mean_dept_salary)) %>%
  slice(1:10)

top_dept_means <- top_dept_means %>%
  select(dept, mean_dept_total_salary = mean_dept_salary)

print(top_dept_means)
```

```
## # A tibble: 10 × 2
##    dept                     mean_dept_total_salary
##    <chr>                                     <dbl>
##  1 Neurosurgery                            380058.
##  2 Provost                                 273790
##  3 Urology                                 216291.
##  4 Orthopaedics                            216205.
##  5 Surgery                                 201917.
##  6 Anesthesiology                          187177.
##  7 Radiation Oncology                      183045.
##  8 Carolina Counts                         182160
##  9 Radiology                               172053.
## 10 Office of the Chancellor                164747.
```

Return a data frame that includes two columns, dept and median_dept_total_salary. The data frame is to order the departments with the highest median total salary appearing first and restrict the departments to the 10 highest paid.

```
top_median_salaries <- dept_summary %>%
  arrange(desc(median_dept_salary)) %>%
  slice(1:10)

top_median_salaries <- top_median_salaries %>%
  select(dept, median_dept_total_salary = median_dept_salary)

print(top_median_salaries)
```

```
## # A tibble: 10 × 2
##    dept                     median_dept_total_salary
##    <chr>                                       <dbl>
##  1 Neurosurgery                               395550
##  2 Provost                                    240080
##  3 Orthopaedics                               240000
##  4 Urology                                    237500
##  5 Anesthesiology                             222645
##  6 Carolina Counts                            182160
##  7 Radiation Oncology                         180000
##  8 Surgery                                    176083
##  9 University Ombuds Office                   157127
## 10 Ath Basketball Office                      150000
```

Why do these lists differ? If you were asked for the top 10 best paid departments at the state university which summary would you choose and why?

These lists differ because they measure different aspects of the same data set, the original data frame filters out the a summary of each department data; there is no explicit comparison or ranking made. For the next two data frames, they each measure a statistic from the fulltime data frame; one measures mean salary and the other measures median salary, both ranked by top 10 highest salaries. I would choose to look at the data from looking at the top 10 median salaries because looking at the mean measure, there is a possibility of outliers that may change the data and be unrepresentative of the actual total salaries.

Create a boxplot of the total salary for all the employees in the 10 departments with the highest mean salaries.

```
top_dept_means <- dept_summary %>%
  arrange(desc(mean_dept_salary)) %>%
  slice(1:10)

top_dept_means <- top_dept_means %>%
  select(dept, mean_dept_total_salary = mean_dept_salary)

top_10_departments <- top_dept_means$dept

filter_10 <- fulltime %>%
  filter(dept %in% top_10_departments)

ggplot(filter_10, aes(y = dept, x = total_salary)) +
  geom_boxplot() +
  labs(x = "Total Salary", y = "Department", title = "Total Salary by Top 10 Departments
with Highest Mean Salaries")
```

Question 8

What departments have at least 25 employees?

```
dept_employee_count <- fulltime %>%
  group_by(dept) %>%
  summarise(employees_count = n()) %>%
  filter(employees_count >= 25)

departments_25_employees <- dept_employee_count$dept
print(departments_25_employees)
```

```
##   [1] "AHEC Support-Comm Med Care"     "Academic Advising"
##   [3] "Admissions"                     "Allied Health Sciences"
##   [5] "Anesthesiology"                 "Anthropology"
##   [7] "Art"                            "Arts & Sci Info Services"
##   [9] "Arts & Sciences Dean's Office"  "Asian Studies"
##  [11] "Ath Olympic Sports"             "Ath Outdoor Facility Oper"
##  [13] "Auxil Enterprises-Gen Adm"      "Biochemistry and Biophysics"
##  [15] "Biology"                        "Biostatistics"
##  [17] "Building Services"              "Business Operations"
##  [19] "Campus Health Services"         "Carolina Institute for DD"
##  [21] "Carolina Population Center"     "Carolina Union"
##  [23] "Cell Biology and Physiology"    "Chemistry"
##  [25] "Clinical Affairs"               "Communication Studies"
##  [27] "Comprehensive Cancer Center"    "Computer Science"
##  [29] "Ctr Health Prom Disease Prev"   "Cys Fibrosis/Pulmonary Res"
##  [31] "Dental Ecology Dept"            "Dental Research"
##  [33] "Dermatology"                    "Design and Construction Svcs"
##  [35] "Dramatic Art"                   "Economics"
##  [37] "Emergency Medicine"             "Energy Services"
##  [39] "English & Comp Literature"      "Environment Sciences & Engi"
##  [41] "Environment, Health & Safety"   "Epidemiology"
##  [43] "Exercise & Sport Science"       "FPG Child Development Inst"
##  [45] "Family Medicine"                "Gastroint Biology & Dis Ctr"
##  [47] "Gene Therapy Center"            "Genetics"
##  [49] "Geography"                      "Germanic & Slavic Lang & Lit"
##  [51] "Global Health & Infec Disease"  "Graduate School"
##  [53] "Grounds Services"               "Health Behavior"
##  [55] "Health Policy and Management"   "Health Sciences Library"
##  [57] "Highway Safety Research"        "History"
##  [59] "Housekeeping Services"          "Housing Res Education"
##  [61] "Human Resources"                "Information Technology Svcs."
##  [63] "Institute of Marine Sciences"   "Journalism/Mass Communication"
##  [65] "Kenan-Flagler Business School"  "Laboratory Animal Medicine"
##  [67] "Maternal & Child Health"        "Mathematics"
##  [69] "Medical Education"              "Medicine"
##  [71] "Medicine Administration"        "Microbiology & Immunology"
##  [73] "Morehead Planetarium"           "Music"
##  [75] "NC Botanical Garden"            "Neurology"
##  [77] "Nutrition"                      "Obstetrics and Gynecology"
##  [79] "Office of Sponsored Research"   "Operative Dentistry"
##  [81] "Ophthalmology"                  "Oral Surgery"
##  [83] "Orthodontics"                   "Orthopaedics"
##  [85] "Otolaryngology (Ent)"           "Pathology & Lab Medicine"
##  [87] "Pediatric Dentistry"            "Pediatrics"
##  [89] "Pharmacology"                   "Philosophy"
##  [91] "Physics-Astronomy"              "Political Science"
##  [93] "Prosthodontics"                 "Psychiatry"
##  [95] "Psychology"                     "Public Policy"
##  [97] "Public Safety"                  "Public Safety Trans & Parking"
##  [99] "Radiation Oncology"             "Radiology"
## [101] "Renaissance Computing Inst"     "Romance Languages"
## [103] "Scholarships & Student Aid"     "School of Dentistry"
```

```
## [105] "School of Education"            "School of Government"
## [107] "School of Info & Libr Science"  "School of Law"
## [109] "School of Nursing"              "School of Pharmacy"
## [111] "School of Public Health"        "School of Social Work"
## [113] "Sheps Ctr for Hlth Serv Res"    "Social Medicine"
## [115] "Sociology"                      "Surgery"
## [117] "TEACCH Autism Program"          "TraCS Institute"
## [119] "UNC Global"                     "UNC Inst for the Environment"
## [121] "UNC McAllister Heart Institute" "University Library"
## [123] "University Registrar"           "V Chancellor-Univ Development"
## [125] "VC for Research"                "VC-Comm and Pub Affair"
## [127] "WUNC-FM"                        "Wm&Ida Friday Ctr-Cont Educ"
```

What departments hired the most employees in 2010? List the top 10 hiring departments and the number of hires in each department.

```
salary_data$hiredate <- ymd(salary_data$hiredate)

class(salary_data$hiredate)
```

```
## [1] "Date"
```

```
top_10_hiring_2010 <- salary_data %>%
  filter(between(hiredate, as.Date('2010-01-01'), as.Date('2010-12-31'))) %>%
  group_by(dept) %>%
  summarise(total_hires = n()) %>%
  arrange(desc(total_hires)) %>%
  slice(1:10)

print(top_10_hiring_2010)
```

```
## # A tibble: 10 × 2
##    dept                          total_hires
##    <chr>                               <int>
##  1 Medicine                               37
##  2 Comprehensive Cancer Center            36
##  3 FPG Child Development Inst             19
##  4 Laboratory Animal Medicine             19
##  5 Psychiatry                             19
##  6 School of Pharmacy                     14
##  7 Medicine Administration                13
##  8 Information Technology Svcs.           12
##  9 Kenan-Flagler Business School          12
## 10 Housekeeping Services                  11
```

Create a separate list of the next 10 top hiring departments and the number of hires in each department.

```
salary_data$hiredate <- ymd(salary_data$hiredate)

class(salary_data$hiredate)
```

```
## [1] "Date"
```

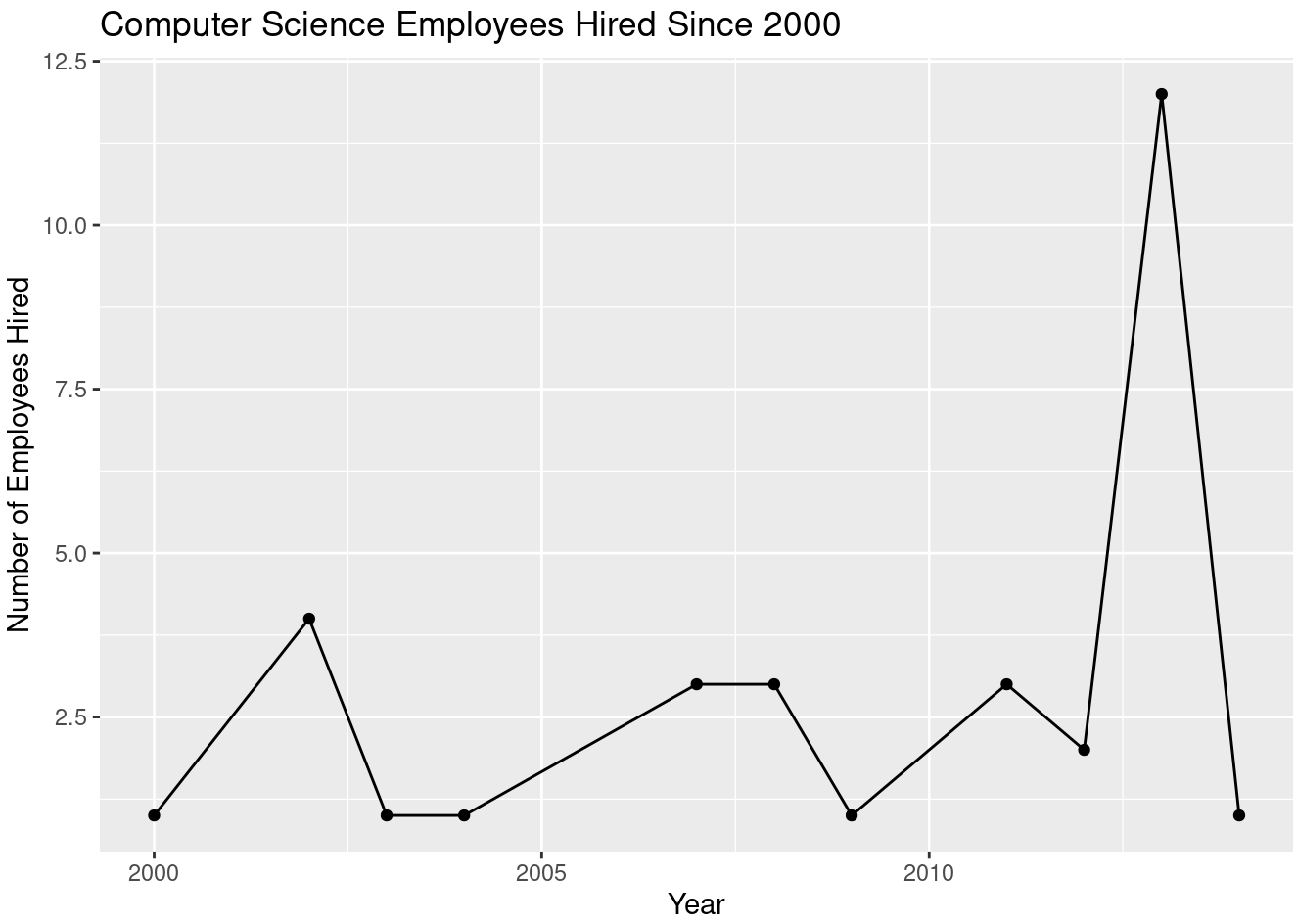
```
next_10_hiring_2010 <- salary_data %>%
  filter(between(hiredate, as.Date('2010-01-01'), as.Date('2010-12-31'))) %>%
  group_by(dept) %>%
  summarise(total_hires = n()) %>%
  arrange(desc(total_hires)) %>%
  slice(11:20)

print(next_10_hiring_2010)
```

```
## # A tibble: 10 × 2
##    dept                          total_hires
##    <chr>                               <int>
##  1 School of Nursing                      11
##  2 University Library                     11
##  3 Epidemiology                           10
##  4 Genetics                               10
##  5 Obstetrics and Gynecology              10
##  6 School of Law                          10
##  7 Surgery                                10
##  8 Campus Health Services                  9
##  9 Journalism/Mass Communication           9
## 10 Pathology & Lab Medicine                8
```

Question 9

Plot number of current employees hired by the Computer Science each year since 2000. The plot is to include points and lines connecting the points

```
salary_data$hiredate <- as.Date(salary_data$hiredate)

cs_hiring_2000 <- salary_data %>%
  filter(dept == "Computer Science", year(hiredate) >= 2000) %>%
  group_by(year = year(hiredate)) %>%
  summarise(employees_hired = n())


ggplot(data = cs_hiring_2000, aes(x = year, y = employees_hired)) +
  geom_point() +
  geom_line() +
  labs(title = "Computer Science Employees Hired Since 2000",
       x = "Year",
       y = "Number of Employees Hired")
```

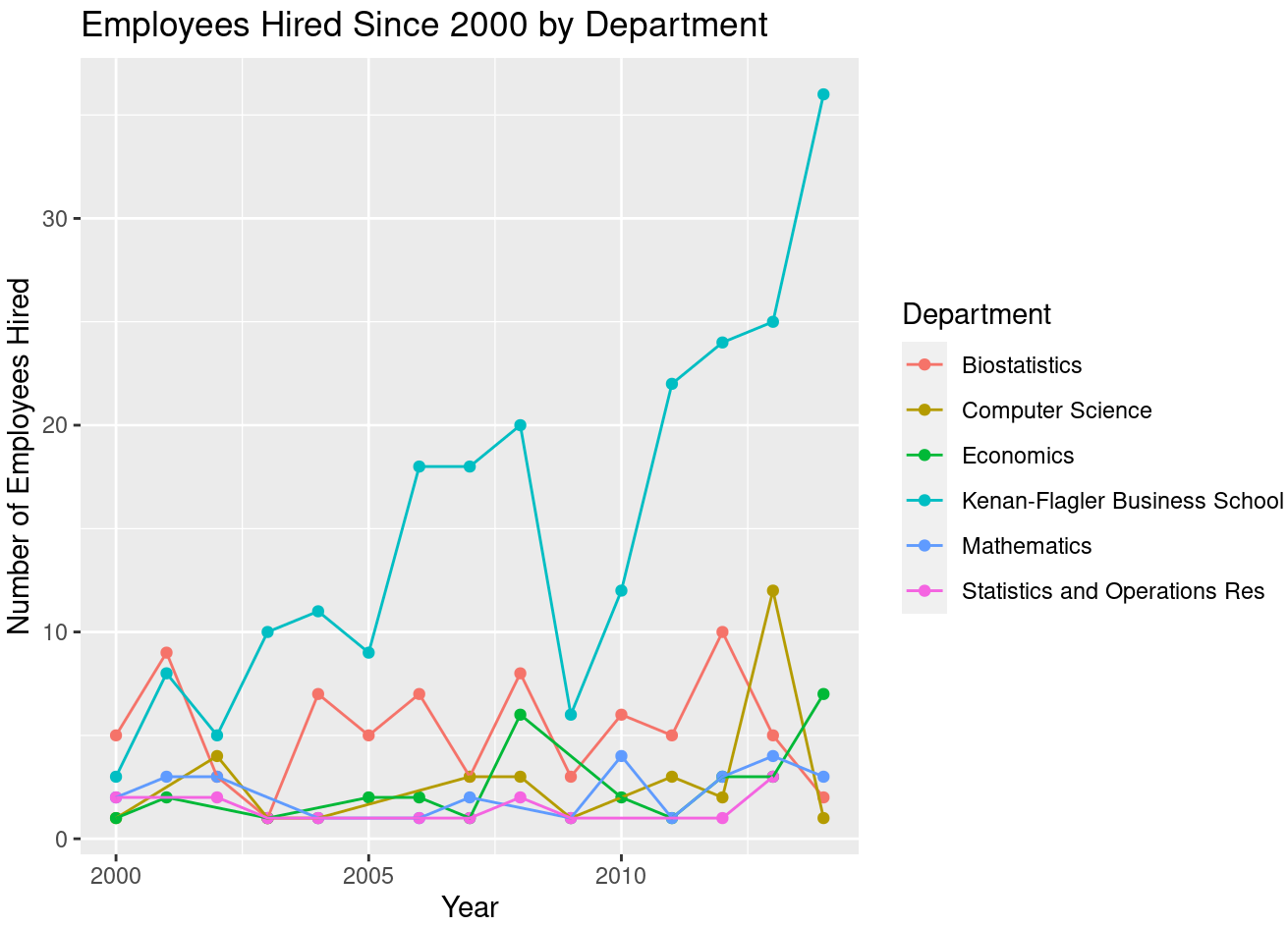


Now add Biostatistics, Economics, Kenan-Flagler Business School, Mathematics, Statistics and Operations Res departments to the above plot.

```
salary_data$hiredate <- as.Date(salary_data$hiredate)

diff_depts <- c("Computer Science", "Biostatistics", "Economics", "Kenan-Flagler Busines
s School", "Mathematics", "Statistics and Operations Res")

dept_hiring_2000 <- salary_data %>%
  filter(dept %in% diff_depts & year(hiredate) >= 2000) %>%
  group_by(dept, year = year(hiredate)) %>%
  summarise(employees_hired = n())
```

```
## `summarise()` has grouped output by 'dept'. You can override using the
## `.groups` argument.
```

```
ggplot(data = dept_hiring_2000, aes(x = year, y = employees_hired, color = dept)) +
  geom_point() +
  geom_line() +
  labs(title = "Employees Hired Since 2000 by Department",
       x = "Year",
       y = "Number of Employees Hired",
       color = "Department")
```



Question 10

```
db = dbConnect(MySQL(),
     user = 'jangc25',
     password = '1749949',
     dbname = 'online_retailer',
     host = 'ballenger.wlu.edu')

knitr::opts_knit$set(sql.max.print = -1)
```

```
SELECT
country,
COUNT(*) "nbr_customers"
FROM customers
GROUP BY country
ORDER BY nbr_customers DESC
```

37 records

country	nbr_customers
United Kingdom	3950
Germany	95

country	nbr_customers
France	87
Spain	29
Belgium	24
Portugal	19
Switzerland	19
Italy	15
Finland	12
Austria	11
Norway	10
Netherlands	9
Channel Islands	9
Australia	9
Sweden	8
Japan	8
Denmark	7
Cyprus	7
Poland	6
USA	4
Greece	4
Unspecified	4
Canada	4
Israel	4
EIRE	3
Malta	2
United Arab Emirates	2
Bahrain	2
Lithuania	1
Lebanon	1
Singapore	1
Saudi Arabia	1

country	nbr_customers
Iceland	1
European Community	1
Czech Republic	1
Brazil	1
RSA	1

Question 11

You are to extract the following data from the online_retailer database for all customers not in the United Kingdom: country, customer_ID, invoice_date, invoice_no (rename invoice_nbr), SKU, description, quantity, and actual_unit_price.

Change the datatype of invoice_date from character to date. The as.Date() function may be handy here. Create a new column for the total sales amount (quantity x price) for each row in the dataframe. Name the new column line_item_total. Create a new column that contains a reformatted the invoice_date. Name the column year_month

```
olr_eu_product_sales <- dbSendQuery(db,
  'SELECT
     country,
     customer_ID,
     invoice_date,
     invoice_no AS invoice_nbr,
     SKU,
     description,
     quantity,
     actual_unit_price
   FROM customers
   JOIN invoices USING(Customer_ID)
   JOIN invoice_products USING(Invoice_No)
   JOIN products USING(SKU)
   WHERE country != "United Kingdom"'
)

olr_eu_product_sales <- dbFetch(olr_eu_product_sales)

olr_eu_product_sales$invoice_date <- as.Date(olr_eu_product_sales$invoice_date)

olr_eu_product_sales$line_item_total <- olr_eu_product_sales$quantity * olr_eu_product_s
ales$actual_unit_price

olr_eu_product_sales$year_month <- format(olr_eu_product_sales$invoice_date, "%Y-%m")
```

We only are interested in data from countries that are currently members of the European Union, see the following dataset, eucountries.csv. You are programmatically use this dataset.

```
eu_countries <- read_csv("data/eucountries.csv")
```

```
## Rows: 27 Columns: 1
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): Country
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
olr_eu_product_sales <- olr_eu_product_sales %>%
  filter(country %in% eu_countries$Country)

cat("Number of observations:", nrow(olr_eu_product_sales), "\n")
```

```
## Number of observations: 195
```

```
cat("Number of variables:", ncol(olr_eu_product_sales), "\n")
```

```
## Number of variables: 10
```

Next you are to create a new dataframe, sales_by_month, that contains the year_month column and the sum of line_item_total column, named total_monthly_sales. Remember to remove any “NA” results. Do not include December 2011 sales in the analysis, as the sales were not collected for a whole month.

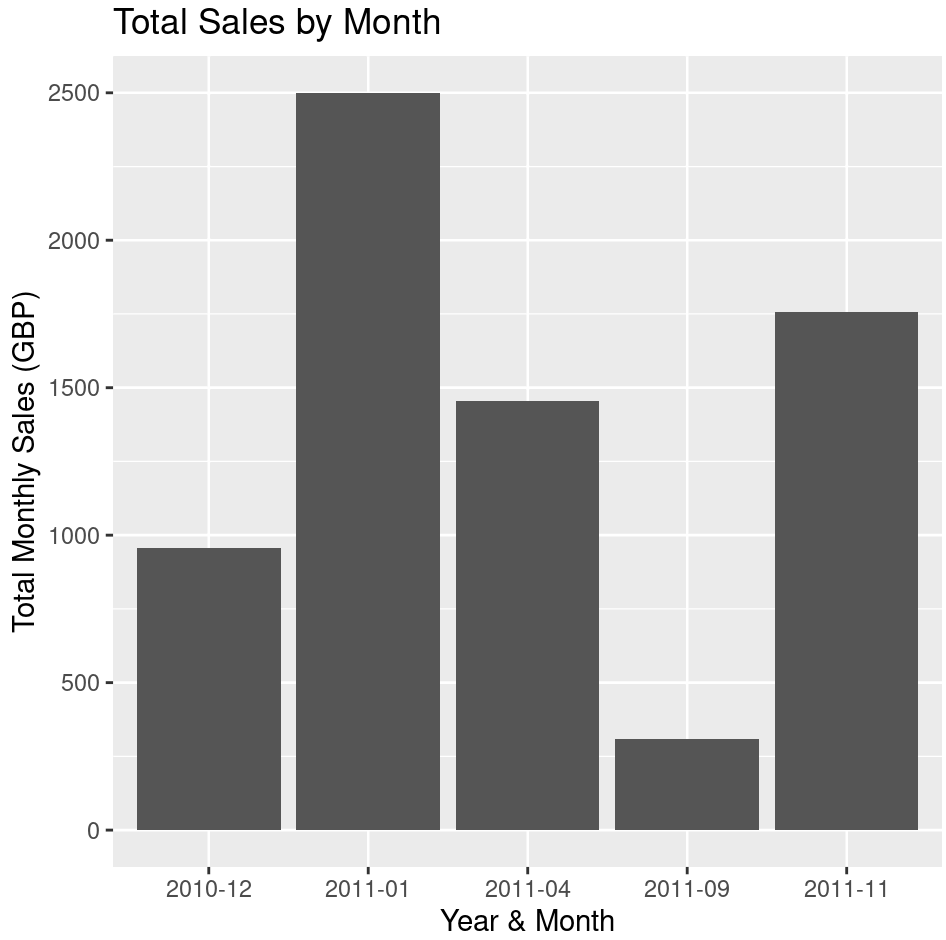
```
sales_by_month <- olr_eu_product_sales %>%
  group_by(year_month) %>%
  summarise(total_monthly_sales = sum(line_item_total, na.rm = TRUE))

sales_by_month <- sales_by_month %>%
  filter(year_month != "2011-12")

print(sales_by_month)
```

```
## # A tibble: 5 × 2
##   year_month total_monthly_sales
##   <chr>                   <dbl>
## 1 2010-12                  955.
## 2 2011-01                 2499.
## 3 2011-04                 1456.
## 4 2011-09                  310
## 5 2011-11                 1758.
```

Question 12

Using the sales_by_month, dataframe to create a “bar chart” type graph with year_month on the x-axis and total_monthly_sales on the y-axis.

```
ggplot(sales_by_month, aes(x = year_month, y = total_monthly_sales)) +
  geom_col() +
  labs(title = "Total Sales by Month",
       x = "Year & Month",
       y = "Total Monthly Sales (GBP)")
```



Project Log

For this project you may also use Google to search for applicable R commands, functions or syntax. If you do use Google you need to cite anything you ultimately use in your project in the Project Log. All other reference material is strictly forbidden

Question 7 ungroup() https://www.statology.org/dplyr-ungroup/ (https://www.statology.org/dplyr-ungroup/)

Question 8 ymd() https://www.rdocumentation.org/packages/lubridate/versions/1.9.3/topics/ymd (https://www.rdocumentation.org/packages/lubridate/versions/1.9.3/topics/ymd)

Question 11- asked Diya Shreenath on advice on going about fixing the read.csv() error -Asked on Wednesday at 10 pm -she said I was missing the /data to read the csv file, and it worked!

Question 11- asked John for help about my file size being too large (it would not knit) -Asked Thursday 10 am -He said that I needed to delete my SQL chunk becuase that was what took up all the file space.

Question 11 cat() https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/cat (https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/cat)

The Pledge

“On my honor, I have neither given nor received any unacknowledged aid on this assignment.” Chaeyon Jang