dplyr & magrittr LAB SOLUTIONS

Paul Intrevado

July 26, 2018

Question 1

The hflights package contains a dataset named hflights, which provides information on 227,496 flights in 2011 leaving from Houston-based airports. Answer the following questions to help you practice your dplyr and magrittr skills.

- How many flights departed per month in total? From IAH per month? From HOU per month?

```
myflights %>% count(Month, Origin) %>% knitr::kable()
```

Month	Origin	n
1	HOU	4270
1	IAH	14640
2	HOU	3884
2	IAH	13244
3	HOU	4544
3	IAH	14926
4	HOU	4420
4	IAH	14173
5	HOU	4533
5	IAH	14639
6	HOU	4499
6	IAH	15101
7	HOU	4519
7	IAH	16029
8	HOU	4505
8	IAH	15671
9	HOU	4186
9	IAH	13879
10	HOU	4405
10	IAH	14291
11	HOU	4212
11	IAH	13809
12	HOU	4322
12	IAH	14795

- What was the airline with the most total departures from IAH? From HOU?

```
myflights %>% group_by(Origin) %>% count(UniqueCarrier, sort = T) %>%
    top_n(1) %>% knitr::kable()
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
## Selecting by n
```

Origin	UniqueCarrier	n
IAH	XE	73053
HOU	WN	45343

- How many flights were cancelled in 2011?

```
myflights %>% summarize(sum(Cancelled))
```

```
##   sum(Cancelled)
## 1           2973
```

- Which airline suffered from the most cancelled flights?

```
myflights %>% group_by(UniqueCarrier) %>% summarise(x = sum(Cancelled)) %>%
    arrange(desc(x)) %>% slice(1)
```

```
## # A tibble: 1 x 2
##   UniqueCarrier     x
##   <chr>         <int>
## 1 XE             1132
```

- Which airline cancelled the most flights *relative* to their total number of flights?

```
myflights %>% group_by(UniqueCarrier) %>% summarise(x = sum(Cancelled),
    y = n()) %>% mutate(prctCancelled = x/y * 100) %>% arrange(desc(prctCancelled)) %>%
    top_n(1)
```

```
## Selecting by prctCancelled
```

```
## # A tibble: 1 x 4
##   UniqueCarrier     x     y prctCancelled
##   <chr>         <int> <int>         <dbl>
## 1 EV               76  2204          3.45
```

- What are the top 3 airlines with the longest mean departure delay?

```
myflights %>% group_by(UniqueCarrier) %>% summarize(meanDepDelay = mean(DepDelay,
    na.rm = T)) %>% arrange(desc(meanDepDelay)) %>% top_n(3) %>%
    knitr::kable()
```

```
## Selecting by meanDepDelay
```

UniqueCarrier	meanDepDelay
WN	13.48824
B6	13.32053
UA	12.91871

UniqueCarrier	meanDepDelay

- Create a table of all airlines describing the mean, median and variance of departure delay, ordered alphabetically by airline? (do this in a single pipe)

```r
myflights %>% select(UniqueCarrier, DepDelay) %>% group_by(UniqueCarrier) %>%
    summarise_all(funs(mean(., na.rm = T), median(., na.rm = T),
        var(., na.rm = T))) %>% arrange(UniqueCarrier) %>% knitr::kable()
```

UniqueCarrier	mean	median	var
AA	6.390144	-2	1250.0659
AS	3.712329	-3	411.7275
B6	13.320532	-2	1837.6027
CO	9.261313	2	670.7362
DL	9.370627	-1	1595.2272
EV	12.482193	-2	1801.8963
F9	5.093637	-2	562.4311
FL	4.716376	-3	1001.6518
MQ	11.071745	-2	1912.1906
OO	8.885482	0	758.3176
UA	12.918707	0	2083.3212
US	1.622926	-4	520.2533
WN	13.488241	4	863.6453
XE	7.713728	-1	789.0647
YV	1.538461	-2	186.3816

- Which airline had the longest mean arrival delay?

```r
myflights %>% group_by(UniqueCarrier) %>% summarise(meanDepDelay = mean(ArrDelay,
    na.rm = T)) %>% arrange(desc(meanDepDelay)) %>% slice(1)
```

```
## # A tibble: 1 x 2
##   UniqueCarrier meanDepDelay
##   <chr>                <dbl>
## 1 UA                    10.5
```

- Which on which day of the week are there the most flights?

```r
myflights %>% count(DayOfWeek, sort = T)
```

```
## # A tibble: 7 x 2
##   DayOfWeek     n
##       <int> <int>
## 1         5 34972
## 2         4 34902
## 3         1 34360
## 4         7 32058
## 5         3 31926
## 6         2 31649
## 7         6 27629
```

- Which carrier has the worst `AirTime` to `Actual Elapsed Time` ratio (the latter of which includes taxiing

```
myflights %>% mutate(x = AirTime/ActualElapsedTime) %>% group_by(UniqueCarrier) %>%
    summarise(y = mean(x, na.rm = T)) %>% arrange(y) %>% slice(1)
```

```
## # A tibble: 1 x 2
##   UniqueCarrier     y
##   <chr>         <dbl>
## 1 AA            0.722
```

- Which flights had a delayed departure but arrived before scheduled time?

```
hflights %>% filter(DepDelay > 0, ArrDelay < 0) %>% glimpse()
```

```
## Observations: 27,712
## Variables: 21
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 20...
## $ Month             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ DayofMonth        <int> 2, 5, 18, 18, 12, 13, 26, 1, 10, 12, 15, 17,...
## $ DayOfWeek         <int> 7, 3, 2, 2, 3, 4, 3, 6, 1, 3, 6, 1, 4, 7, 6,...
## $ DepTime           <int> 1401, 1405, 1408, 721, 2015, 2020, 2009, 163...
## $ ArrTime           <int> 1501, 1507, 1508, 827, 2113, 2116, 2103, 173...
## $ UniqueCarrier     <chr> "AA", "AA", "AA", "AA", "AA", "AA", "AA", "A...
## $ FlightNum         <int> 428, 428, 428, 460, 533, 533, 533, 1121, 112...
## $ TailNum           <chr> "N557AA", "N492AA", "N507AA", "N558AA", "N55...
## $ ActualElapsedTime <int> 60, 62, 60, 66, 58, 56, 54, 65, 61, 68, 64, ...
## $ AirTime           <int> 45, 44, 42, 46, 39, 44, 39, 37, 41, 44, 48, ...
## $ ArrDelay          <int> -9, -3, -2, -8, -7, -4, -17, -9, -5, -6, -9,...
## $ DepDelay          <int> 1, 5, 8, 1, 10, 15, 4, 1, 9, 1, 2, 2, 4, 5, ...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "I...
## $ Dest              <chr> "DFW", "DFW", "DFW", "DFW", "DFW", "DFW", "D...
## $ Distance          <int> 224, 224, 224, 224, 224, 224, 224, 224, 224,...
## $ TaxiIn            <int> 6, 9, 7, 7, 9, 4, 9, 16, 8, 5, 5, 10, 10, 9,...
## $ TaxiOut           <int> 9, 9, 11, 13, 10, 8, 6, 12, 12, 19, 11, 11, ...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ CancellationCode  <chr> "", "", "", "", "", "", "", "", "", "", "", ...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

- Create a new hflights1 dataframe with an additional variable delay_percent to the dataset.

```
hflights1 <- hflights %>% mutate(delay_percent = (ArrDelay -
    DepDelay)/DepDelay * 100)
```

- Use `airlines` to rename the carriers

```
airlines <- c(AA = "American", AS = "Alaska", B6 = "JetBlue",
    CO = "Continental", DL = "Delta", OO = "SkyWest", UA = "United",
    US = "US_Airways", WN = "Southwest", EV = "Atlantic_Southeast",
    F9 = "Frontier", FL = "AirTran", MQ = "American_Eagle", XE = "ExpressJet",
    YV = "Mesa")
```

```r
hflights$UniqueCarrier <- airlines[hflights$UniqueCarrier]
```

- Find the flights flown by one of JetBlue, American_Eagle, or Continental

```r
hflights %>% filter(UniqueCarrier %in% c("JetBlue", "American_Eagle",
    "Continental")) %>% glimpse()
```

```
## Observations: 75,375
## Variables: 21
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 20...
## $ Month             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ DayofMonth        <int> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 7, 7, 8, 9,...
## $ DayOfWeek         <int> 6, 6, 7, 7, 1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 7,...
## $ DepTime           <int> 654, 1639, 703, 1604, 659, 1801, 654, 1608, ...
## $ ArrTime           <int> 1124, 2110, 1113, 2040, 1100, 2200, 1103, 20...
## $ UniqueCarrier     <chr> "JetBlue", "JetBlue", "JetBlue", "JetBlue", ...
## $ FlightNum         <int> 620, 622, 620, 622, 620, 622, 620, 622, 620,...
## $ TailNum           <chr> "N324JB", "N324JB", "N324JB", "N324JB", "N22...
## $ ActualElapsedTime <int> 210, 211, 190, 216, 181, 179, 189, 206, 183,...
## $ AirTime           <int> 181, 188, 172, 176, 166, 165, 168, 175, 167,...
## $ ArrDelay          <int> 5, 61, -6, 31, -19, 111, -16, 25, -14, -6, -...
## $ DepDelay          <int> -6, 54, 3, 19, -1, 136, -6, 23, 0, 9, -3, -6...
## $ Origin            <chr> "HOU", "HOU", "HOU", "HOU", "HOU", "HOU", "H...
## $ Dest              <chr> "JFK", "JFK", "JFK", "JFK", "JFK", "JFK", "J...
## $ Distance          <int> 1428, 1428, 1428, 1428, 1428, 1428, 1428, 14...
## $ TaxiIn            <int> 6, 12, 6, 9, 3, 5, 9, 8, 4, 14, 7, 6, 9, 9, ...
## $ TaxiOut           <int> 23, 11, 12, 31, 12, 9, 12, 23, 12, 10, 9, 25...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ CancellationCode  <chr> "", "", "", "", "", "", "", "", "", "", "", ...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

- Which flights had taxiing time that was greater than flying time? (where taxiing: `TaxinIn` + `TaxiOut`)

```r
hflights %>% filter((TaxiIn + TaxiOut) > AirTime) %>% glimpse()
```

```
## Observations: 1,389
## Variables: 21
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 20...
## $ Month             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ DayofMonth        <int> 24, 30, 24, 10, 31, 31, 31, 31, 30, 30, 30, ...
## $ DayOfWeek         <int> 1, 7, 1, 1, 1, 1, 1, 1, 7, 7, 7, 7, 7, 7, 7,...
## $ DepTime           <int> 731, 1959, 1621, 941, 1301, 2113, 1434, 900,...
## $ ArrTime           <int> 904, 2132, 1749, 1113, 1356, 2215, 1539, 100...
## $ UniqueCarrier     <chr> "American", "American", "American", "America...
## $ FlightNum         <int> 460, 533, 1121, 1436, 241, 1533, 1541, 1583,...
## $ TailNum           <chr> "N545AA", "N455AA", "N484AA", "N591AA", "N14...
## $ ActualElapsedTime <int> 93, 93, 88, 92, 55, 62, 65, 66, 64, 84, 80, ...
## $ AirTime           <int> 42, 43, 43, 45, 27, 30, 30, 32, 31, 40, 37, ...
## $ ArrDelay          <int> 29, 12, 4, 48, -2, 20, 15, 10, 10, 54, 16, 1...
## $ DepDelay          <int> 11, -6, -9, 31, -4, 13, 4, 0, -1, 39, 2, -4,...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "I...
## $ Dest              <chr> "DFW", "DFW", "DFW", "DFW", "AUS", "AUS", "A...
## $ Distance          <int> 224, 224, 224, 224, 140, 140, 140, 140, 140,...
## $ TaxiIn            <int> 14, 10, 10, 27, 5, 7, 5, 5, 6, 10, 6, 4, 6, ...
## $ TaxiOut           <int> 37, 40, 35, 20, 23, 25, 30, 29, 27, 34, 37, ...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ CancellationCode  <chr> "", "", "", "", "", "", "", "", "", "", "", ...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

- Find all the flights that were cancelled after being delayed

```
hflights %>% filter(DepDelay > 0, Cancelled == 1) %>% glimpse()
```

```
## Observations: 40
## Variables: 21
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 20...
## $ Month             <int> 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 4, 4, 4,...
## $ DayofMonth        <int> 26, 11, 19, 7, 4, 8, 2, 9, 1, 31, 4, 8, 21, ...
## $ DayOfWeek         <int> 3, 2, 3, 5, 5, 2, 3, 3, 2, 4, 1, 5, 4, 1, 1,...
## $ DepTime           <int> 1926, 1100, 1811, 2028, 1638, 1057, 802, 904...
## $ ArrTime           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ UniqueCarrier     <chr> "Continental", "US_Airways", "ExpressJet", "...
## $ FlightNum         <int> 310, 944, 2376, 3050, 1121, 408, 2189, 2605,...
## $ TailNum           <chr> "N77865", "N452UW", "N15932", "N15912", "N53...
## $ ActualElapsedTime <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ AirTime           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ ArrDelay          <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ DepDelay          <int> 26, 135, 6, 73, 8, 187, 2, 4, 28, 156, 42, 5...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "I...
## $ Dest              <chr> "EWR", "CLT", "ICT", "JAX", "DFW", "EWR", "D...
## $ Distance          <int> 1400, 913, 542, 817, 224, 1400, 217, 217, 68...
## $ TaxiIn            <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ TaxiOut           <int> NA, NA, NA, 19, 19, NA, NA, NA, 19, NA, NA, ...
## $ Cancelled         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ CancellationCode  <chr> "B", "B", "B", "A", "A", "A", "B", "B", "A",...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

- Display all the flights leaving IAH before 10 am and arrange according to decreasing `AirTime`

```
hflights %>% filter(Origin == "IAH", DepTime < 800) %>% arrange(desc(AirTime)) %>%
    glimpse()
```

```
## Observations: 17,835
## Variables: 21
## $ Year              <int> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 20...
## $ Month             <int> 8, 2, 12, 3, 3, 12, 11, 3, 5, 11, 10, 12, 12...
## $ DayofMonth        <int> 1, 28, 31, 6, 31, 29, 14, 10, 20, 11, 17, 30...
## $ DayOfWeek         <int> 1, 1, 6, 7, 4, 4, 1, 4, 5, 5, 1, 5, 3, 1, 1,...
## $ DepTime           <int> 156, 752, 733, 747, 750, 731, 733, 748, 744,...
## $ ArrTime           <int> 452, 1100, 1048, 1052, 1100, 1122, 1032, 104...
## $ UniqueCarrier     <chr> "Continental", "Continental", "Continental",...
## $ FlightNum         <int> 1, 167, 1551, 167, 167, 1551, 167, 167, 167,...
## $ TailNum           <chr> "N69063", "N37293", "N17244", "N73283", "N18...
## $ ActualElapsedTime <int> 476, 308, 315, 305, 310, 351, 299, 295, 294,...
## $ AirTime           <int> 461, 286, 286, 282, 281, 281, 278, 276, 276,...
## $ ArrDelay          <int> 957, 21, 21, 20, 23, 55, 3, 6, 21, 34, 27, -...
## $ DepDelay          <int> 981, 2, 7, -3, 0, 6, 3, -2, -1, -1, 4, 2, 2,...
## $ Origin            <chr> "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", "I...
## $ Dest              <chr> "HNL", "SEA", "SEA", "SEA", "SEA", "SEA", "S...
## $ Distance          <int> 3904, 1874, 1874, 1874, 1874, 1874, 1874, 18...
## $ TaxiIn            <int> 5, 6, 7, 7, 8, 4, 4, 6, 5, 7, 5, 5, 7, 5, 6,...
## $ TaxiOut           <int> 10, 16, 22, 16, 21, 66, 17, 13, 13, 20, 26, ...
## $ Cancelled         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ CancellationCode  <chr> "", "", "", "", "", "", "", "", "", "", "", ...
## $ Diverted          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

Question 2

The pokemon data set contains information on (all?) Pokemon. Answer the following questions to help you practice your dplyr and magrittr skills.

- How many Pokemon are considered Legendary?

```
pokemon %>% filter(Legendary == "True") %>% summarise(n())
```

```
## # A tibble: 1 x 1
##   `n()`
##   <int>
## 1    65
```

- List the top five Pokeman, based on Total, whose Type 1 is either Grass or Fire.

```
pokemon %>% filter(`Type 1` == "Grass" | `Type 1` == "Fire") %>%
    group_by(`Type 1`) %>% top_n(5, Total)
```

```
## # A tibble: 11 x 13
## # Groups:   Type 1 [2]
##      `#` Name  `Type 1` `Type 2` Total    HP Attack Defense `Sp. Atk`
##    <int> <chr> <chr>    <chr>    <int> <int>  <int>   <int>     <int>
##  1     3 Venu~ Grass    Poison     625    80    100     123       122
##  2     6 Char~ Fire     Dragon     634    78    130     111       130
##  3     6 Char~ Fire     Flying     634    78    104      78       159
##  4   250 Ho-oh Fire     Flying     680   106    130      90       110
##  5   254 Scep~ Grass    Dragon     630    70    110      75       145
##  6   257 Blaz~ Fire     Fighting   630    80    160      80       130
##  7   460 Abom~ Grass    Ice        594    90    132     105       132
##  8   485 Heat~ Fire     Steel      600    91     90     106       130
##  9   492 Shay~ Grass    <NA>       600   100    100     100       100
## 10   492 Shay~ Grass    Flying     600   100    103      75       120
## 11   721 Volc~ Fire     Water      600    80    110     120       130
## # ... with 4 more variables: `Sp. Def` <int>, Speed <int>,
## #   Generation <int>, Legendary <chr>
```

- What are the mean and standard deviation of HP for each Generation of Pokemon?

```
pokemon %>% group_by(Generation) %>% summarize(myMean = mean(HP),
    mySTD = sd(HP))
```

```
## # A tibble: 6 x 3
##   Generation myMean mySTD
##        <int>  <dbl> <dbl>
## 1          1   65.8  28.2
## 2          2   71.2  30.6
## 3          3   66.5  24.1
## 4          4   73.1  25.1
## 5          5   71.8  22.4
## 6          6   68.3  20.9
```

- A Coefficient of Variation (CoV) is defined as the standard deviation divided by the mean ($\frac{s}{\bar{x}}$). Which `Generation` of Pokemon has the **lowest** Cov for `Attack`?

```
pokemon %>% group_by(Generation) %>% summarize(CoV = sd(HP)/mean(HP)) %>%
    arrange(CoV)
```

```
## # A tibble: 6 x 2
##   Generation   CoV
##        <int> <dbl>
## 1          6 0.306
## 2          5 0.312
## 3          4 0.344
## 4          3 0.362
## 5          1 0.428
## 6          2 0.430
```

- Based on their `Type 2` characteristic, what are the Pokeman with the highest and lowest `Speed`?

```
pokemon %>% top_n(1, Speed)

group_by(`Type 1`) %>% # arrange(desc(Speed)) %>%
top_n(1, Speed) %>% arrange(`Type 1`)
```

Question 3

Import `uncSalaries.csv`, data on the salaries of the University of North Carolina's employees.

- What is the mean salary in the Neurosurgery department?

```
unc %>% filter(dept == "Neurosurgery") %>% summarise(meanSal = mean(totalsal,
    na.rm = T))
```

```
## # A tibble: 1 x 1
##   meanSal
##     <dbl>
## 1 380058.
```

- Return a data frame with employee's in the Neurosurgery department making more than $500,000. Why might these professors be so well paid?

```
unc %>% filter(dept == "Neurosurgery", totalsal > 5e+05)
```

```
## # A tibble: 6 x 14
##   name  campus dept  position exempt2 employed hiredate   fte status
##   <chr> <chr>  <chr> <chr>    <chr>      <int>    <int> <dbl> <chr>
## 1 CAMP~ UNC-CH Neur~ Adjunct~ Exempt        12 20140731     1 Fixed~
## 2 CARS~ UNC-CH Neur~ Clinica~ Exempt        12 20090430     1 Fixed~
## 3 EWEN~ UNC-CH Neur~ DIRECTOR Exempt        12 19970731     1 Conti~
## 4 JAUF~ UNC-CH Neur~ Clinica~ Exempt        12 20080930     1 Fixed~
## 5 KILP~ UNC-CH Neur~ Clinica~ Exempt        12 20130930     1 Fixed~
## 6 WADO~ UNC-CH Neur~ Clinica~ Exempt        12 20080930     1 Fixed~
## # ... with 5 more variables: stservyr <int>, statesal <int>,
## #   nonstsal <int>, totalsal <int>, age <int>
```

- What is the total amount that full time Dermatology employees get paid

```
unc %>% filter(dept == "Dermatology", fte == 1) %>% summarise(sum(totalsal))
```

```
## # A tibble: 1 x 1
##   `sum(totalsal)`
##             <int>
## 1         5272098
```

- Create a data frame called `radio_dept` whose rows are the employees from the Radiology department.
  - include only the following columns: `name`, `position`, `age`, `nonstsal`, `totalsal`.
  - order the employees by salary

```
unc %>% filter(dept == "Radiology") %>% select(name, position,
    age, nonstsal, totalsal) %>% arrange(desc(totalsal))
```

```
## # A tibble: 88 x 5
##    name               position                       age nonstsal totalsal
##    <chr>              <chr>                        <int>    <int>    <int>
##  1 MAURO, MATTHEW A   DIRECTOR                        63   614176   614176
##  2 LEE, JOSEPH K      Professor                       67   375000   375000
##  3 BURKE, CHARLES T   Clinical Associate Professor    44   365000   365000
##  4 MOLINA, PAUL L     Professor                       56   334255   350000
##  5 STAVAS, JOSEPH M   Clinical Professor              59   345000   345000
##  6 DIXON, ROBERT G    Clinical Associate Professor    55   335000   335000
##  7 CASTILLO, MAURICIO Professor                       55   316255   332000
##  8 SEMELKA, RICHARD C Professor                       54   306255   322000
##  9 SMITH, J K         Professor with Tenure           52   292187   310000
## 10 FIELDING, JULIA R  Associate Professor             53   294005   309750
## # ... with 78 more rows
```

- 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 (using totalsal for salary).

```
dept_summary <- unc %>% group_by(dept) %>% summarise(deptSize = n(),
    medSal = median(totalsal, na.rm = T), maxSal = max(totalsal,
        na.rm = T))
```

- Order the departments by highest mean salary and print the 10 highest paid departments.

```
unc %>% group_by(dept) %>% summarise(meanSal = mean(totalsal,
    na.rm = T)) %>% arrange(desc(meanSal)) %>% top_n(10, meanSal)
```

```
## # A tibble: 10 x 2
##    dept                    meanSal
##    <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.
```

- Order the departments by highest median salary and print the 10 highest paid departments.

```
unc %>% group_by(dept) %>% summarise(medSal = median(totalsal,
    na.rm = T)) %>% arrange(desc(medSal)) %>% top_n(10, medSal)
```

```
## # A tibble: 10 x 2
##    dept                     medSal
##    <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 UNC which summary would you choose and why?