MA342_Project_Analysis.R

crane

2023-11-27

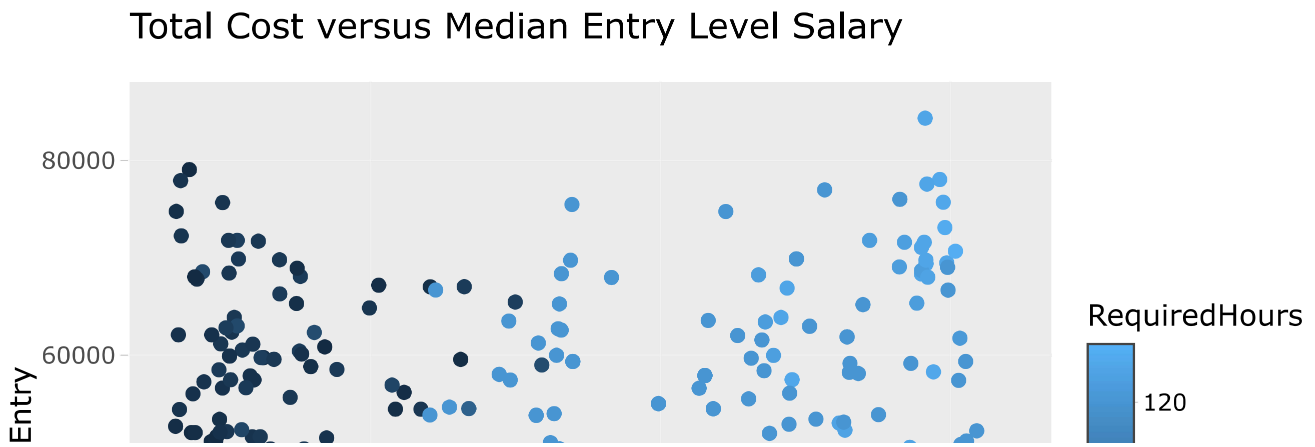
```
options(scipen=999, warn = -1)
library(tidyverse)
```

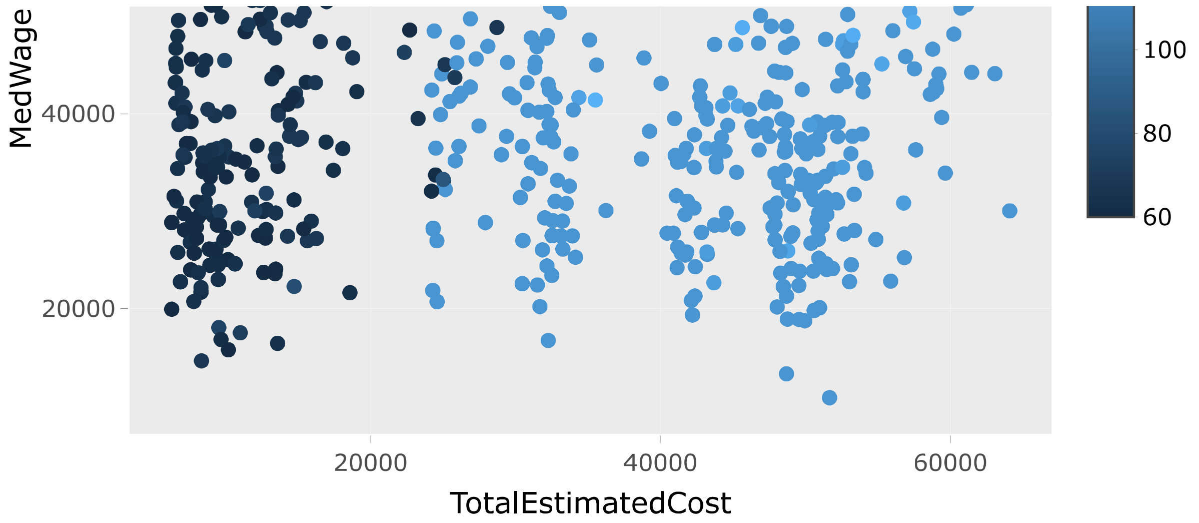
```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.3     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(plotly)
```

```
##
## Attaching package: 'plotly'
##
## The following object is masked from 'package:ggplot2':
##
##     last_plot
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following object is masked from 'package:graphics':
##
##     layout
```

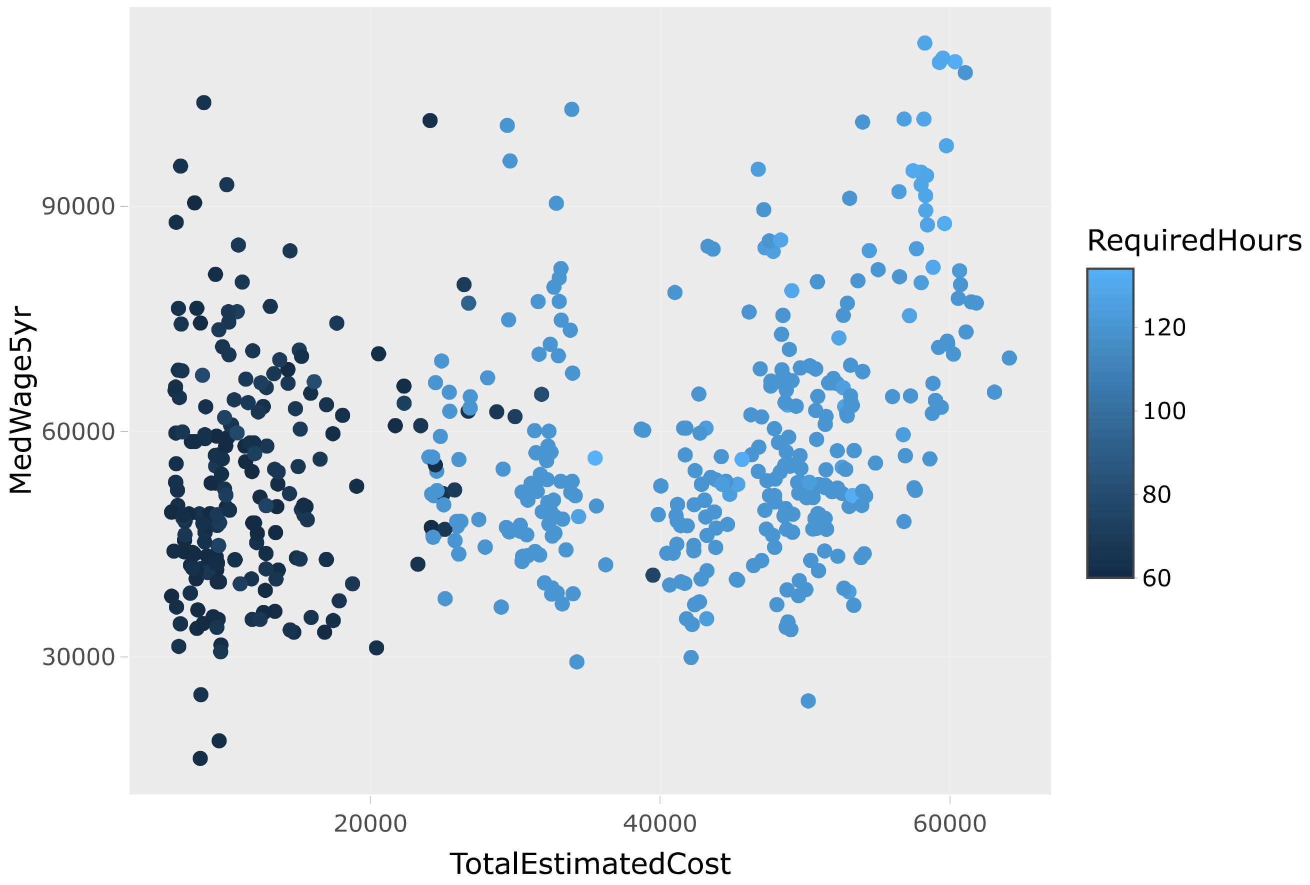
```
courses <- as_tibble(read.csv("C:/Users/crane/Documents/MA342/Project/courses.csv"))
institutions <- read.csv("C:/Users/crane/Documents/MA342/Project/Institutions.csv")


#remove special characters in tuition, fees, placement, and wage columns
#forcing  columns as numeric replaces the ^ and - characters present in data set to NA, which is
what we want
courses <- courses %>%
  mutate(inst_id = as.numeric(inst_id)) %>%
  mutate(AnnualTuition = as.numeric(str_replace(AnnualTuition, "[$%]", ""))) %>%
  mutate(AnnualFees = as.numeric(str_replace(AnnualFees, "[$%]", ""))) %>%
  mutate(Placement = as.numeric(str_replace(Placement, "[$%]", ""))) %>%
  mutate(MedWageEntry = as.numeric(str_replace(MedWageEntry, "[$%]", ""))) %>%
  mutate(MedWage5yr = as.numeric(str_replace(MedWage5yr, "[$%]", "")))

#add columns for total cost per year
courses <- courses %>%
  mutate(AnnualCost = AnnualTuition + AnnualFees) %>%
  mutate(TotalEstimatedCost = (RequiredHours/30)*AnnualCost) %>%
  mutate(Med5YrValue = MedWage5yr/TotalEstimatedCost) %>%
  mutate(MedEntryValue = MedWageEntry/TotalEstimatedCost)


#_______________________________________________________________________________________________
___
#plot showing cost of degree vs median entry wage  of all degrees, color by required hours
#these two plots best answer the question, what degree from KHE schools have the best value
#both of these plots are interactive, meaning they show additional information when you
#hover over one of the data points but must be viewed as an html file
plt1 <- courses %>%
  left_join(institutions, by = "inst_id") %>%
  filter(TotalEstimatedCost < 75000) %>%
  ggplot() +
  geom_point(mapping = aes(x = TotalEstimatedCost, y = MedWageEntry, color = RequiredHours, text
= paste0(DegreeTitle, " ", InstitutionName))) +
  ggtitle("Total Cost versus Median Entry Level Salary")

ggplotly(plt1)
```



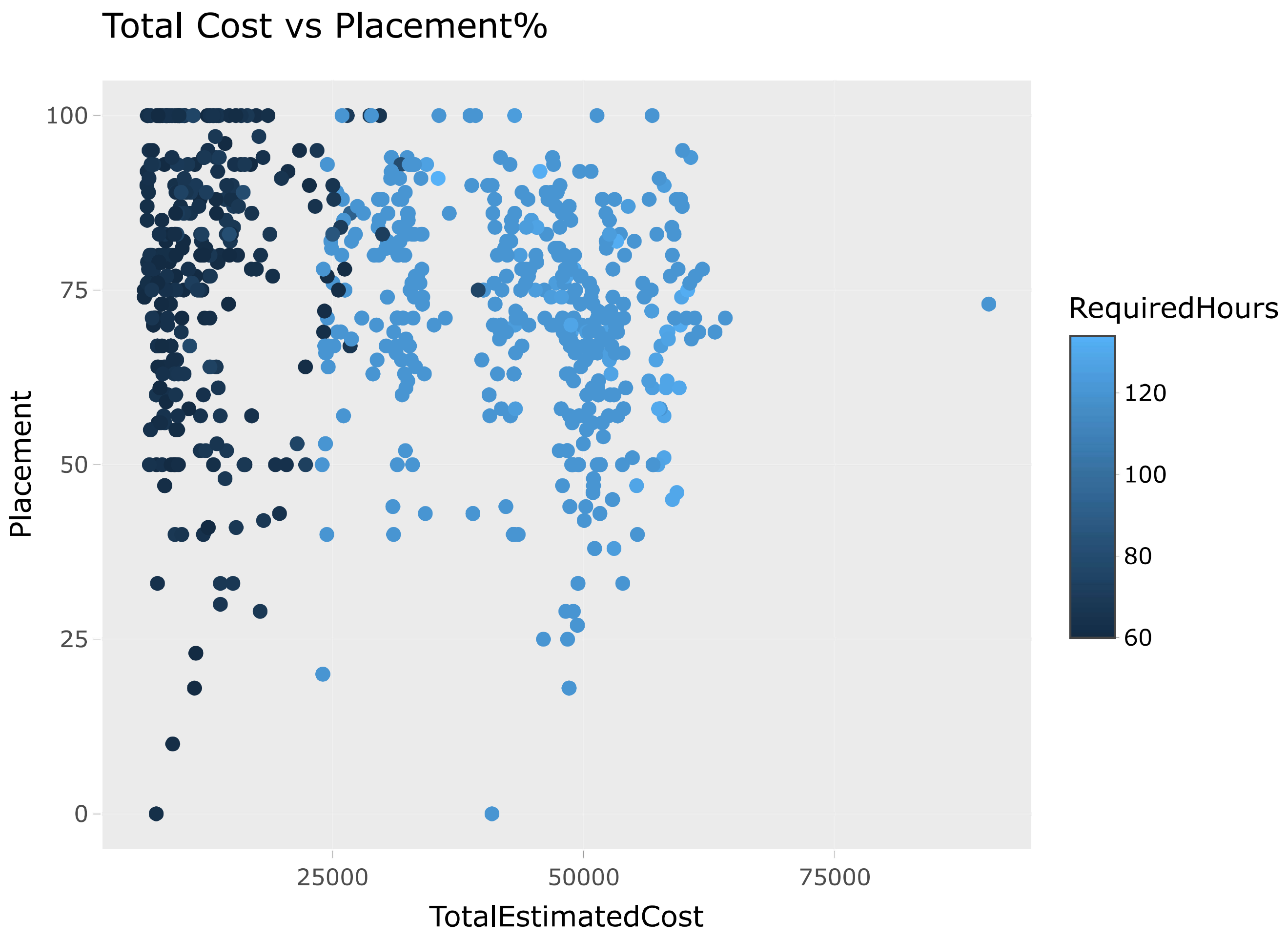


```
plt2 <- courses %>%
  left_join(institutions, by = "inst_id") %>%
  filter(TotalEstimatedCost < 75000) %>%
  ggplot() +
  geom_point(mapping = aes(x = TotalEstimatedCost, y = MedWage5yr, color = RequiredHours, text =
paste0(DegreeTitle, " ", InstitutionName))) +
  ggtitle("Total Cost versus Median Salary After 5 years")

ggplotly(plt2)
```

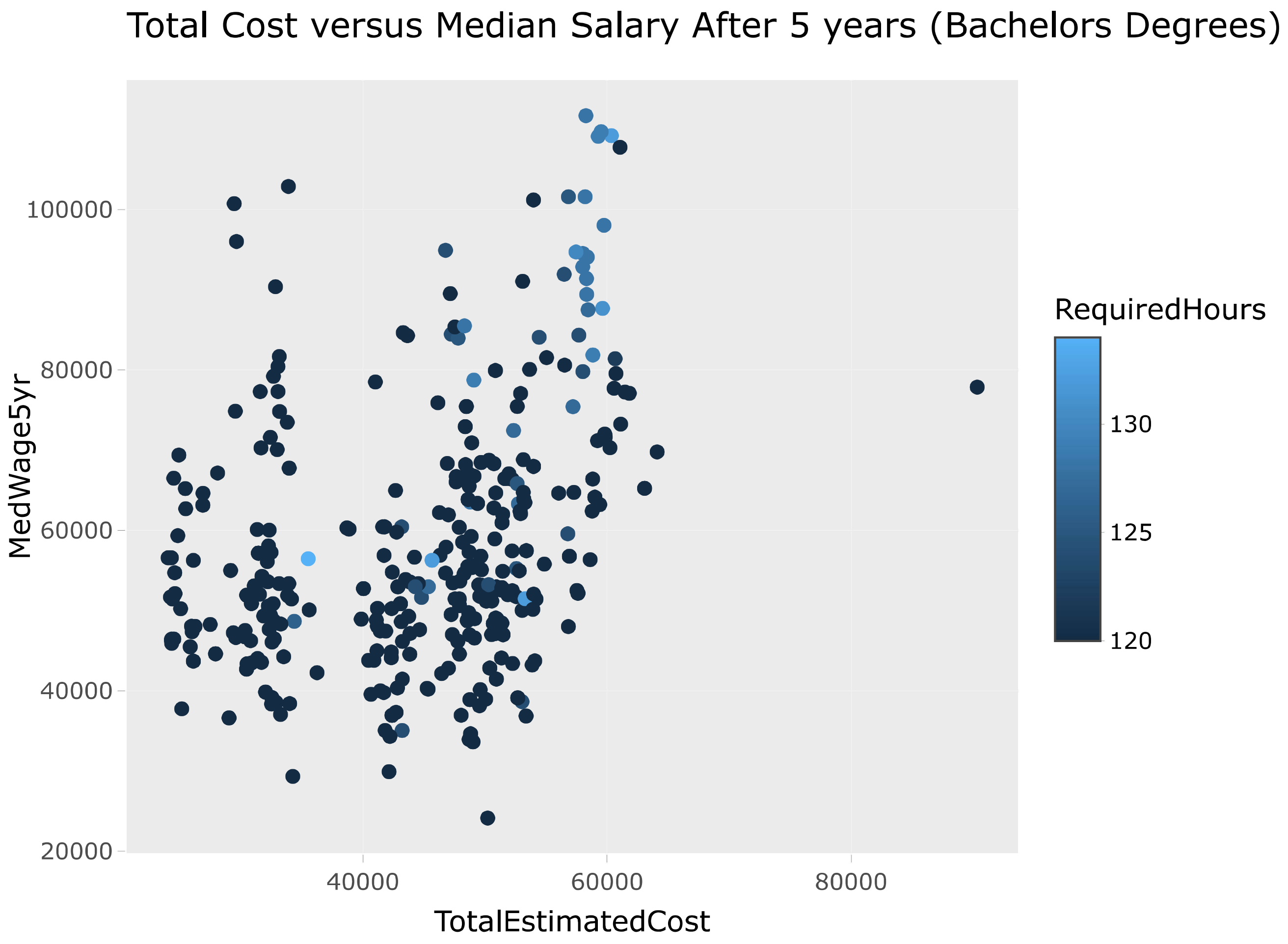
Total Cost versus Median Salary After 5 years



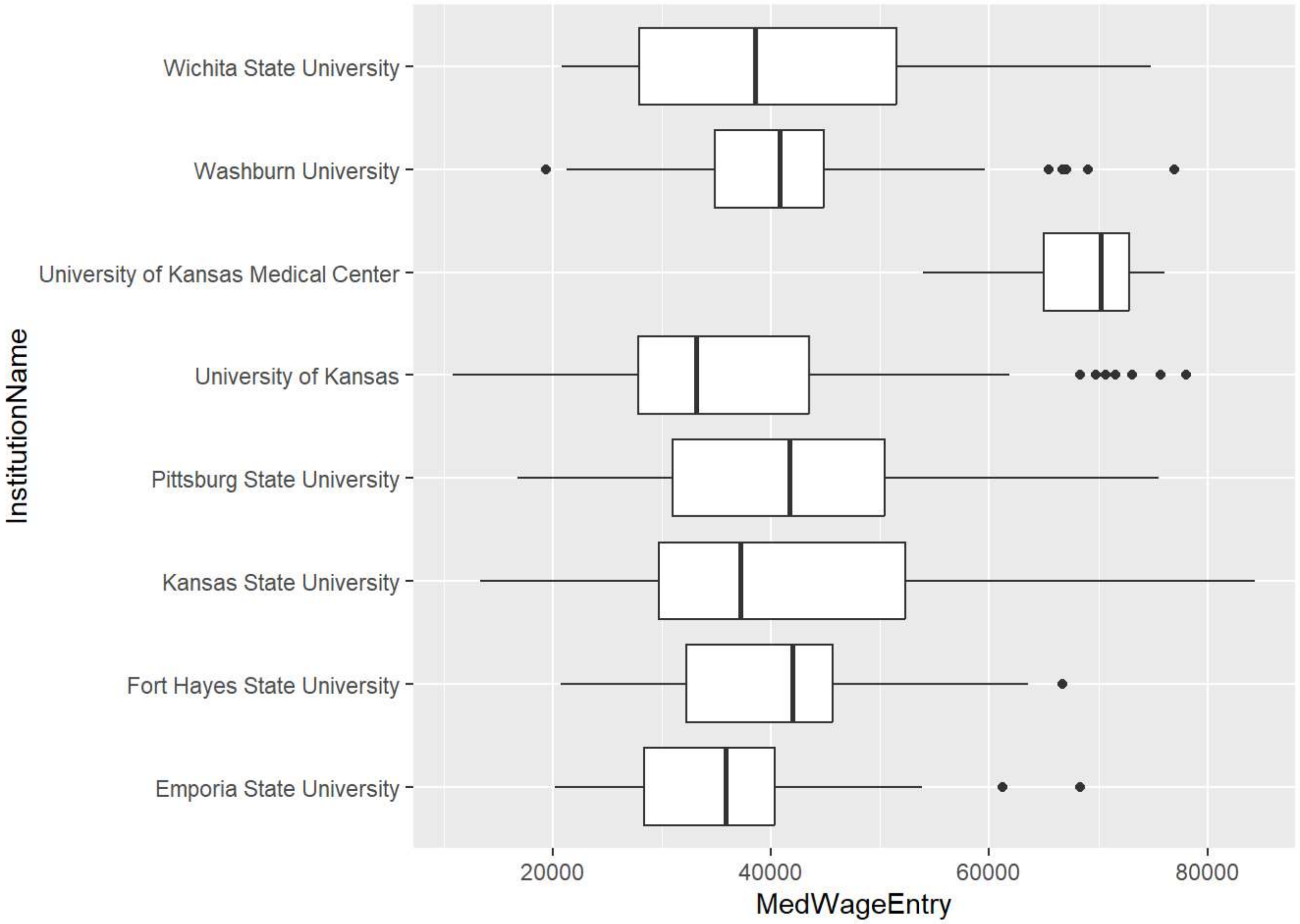
```
plt3 <- courses %>%
  left_join(institutions, by = "inst_id") %>%
  ggplot()+
  geom_point(mapping = aes(x = TotalEstimatedCost, y = Placement, color = RequiredHours, text =
paste0(DegreeTitle, " ", InstitutionName))) +
  ggtitle("Total Cost vs Placement%")

 ggplotly(plt3)
```



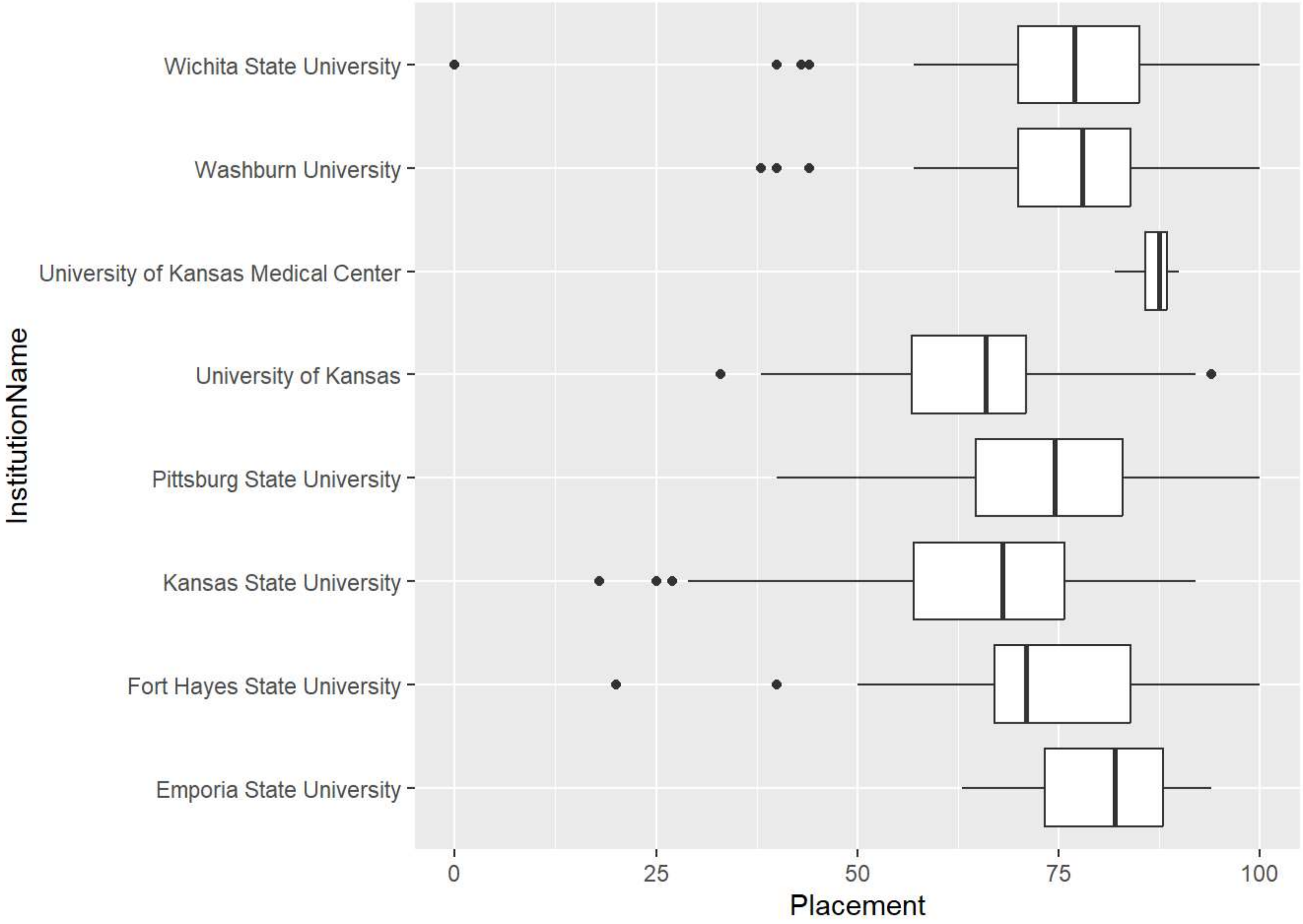
```
#Using the results of these 3 charts, we can make a determination that associates degrees
#have the highest value in terms of cost vs wage and placement percentages

#its also interesting to note that the points are in 3 distinct groups based on cost
#1) two year colleges <20000
#2) Fort Hayes, Pittsburg, Emporia >20000 and < 40000
#3) KU, KSU, WSU, WU >40000

#obviously an associates degree isn't not for everyone, so below is a graph of just the bachelors degrees
plt4 <- courses %>%
  left_join(institutions, by = "inst_id") %>%
  filter(RequiredHours >= 120) %>%
  ggplot() +
  geom_point(mapping = aes(x = TotalEstimatedCost, y = MedWage5yr, color = RequiredHours, text = paste0(DegreeTitle, " ", InstitutionName))) +
  ggtitle("Total Cost versus Median Salary After 5 years (Bachelors Degrees)")

ggplotly(plt4)
```



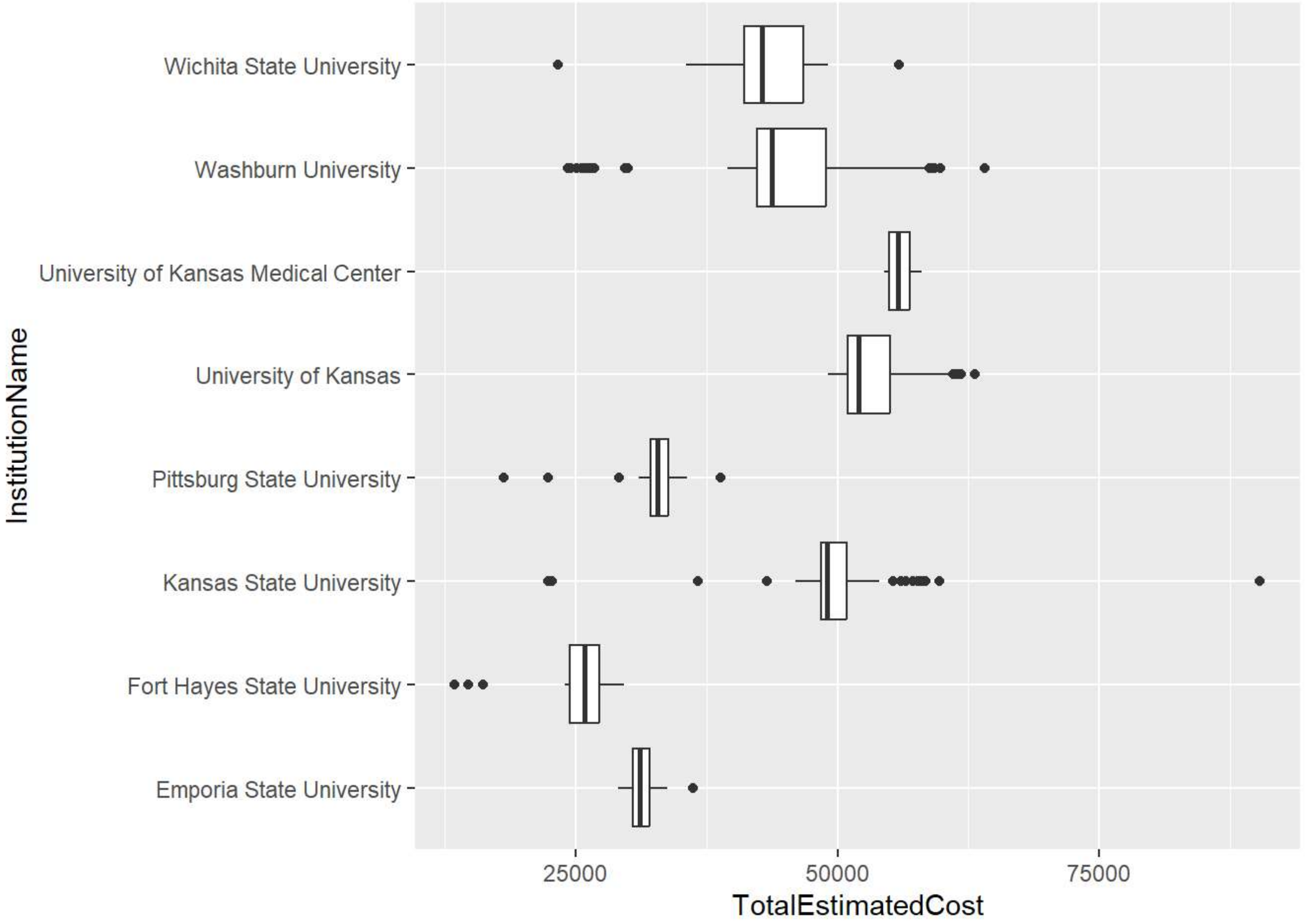
```
#______________________________________________________________________________________________
____

#plot showing distribution of wages by university
courses %>%
  left_join(institutions, by = "inst_id") %>%
  filter(inst_type %in% c("State University", "Municipal University")) %>%
  ggplot() +
  geom_boxplot(mapping = aes(x = InstitutionName, y = MedWageEntry)) +
  coord_flip()
```



```
#placement by university
courses %>%
  left_join(institutions, by = "inst_id") %>%
  filter(inst_type %in% c("State University", "Municipal University")) %>%
  ggplot() +
  geom_boxplot(mapping = aes(x = InstitutionName, y = Placement)) +
  coord_flip()
```



```
#university by costs
courses %>%
  left_join(institutions, by = "inst_id") %>%
  filter(inst_type %in% c("State University", "Municipal University")) %>%
  ggplot() +
  geom_boxplot(mapping = aes(x = InstitutionName, y = TotalEstimatedCost)) +
  coord_flip()
```



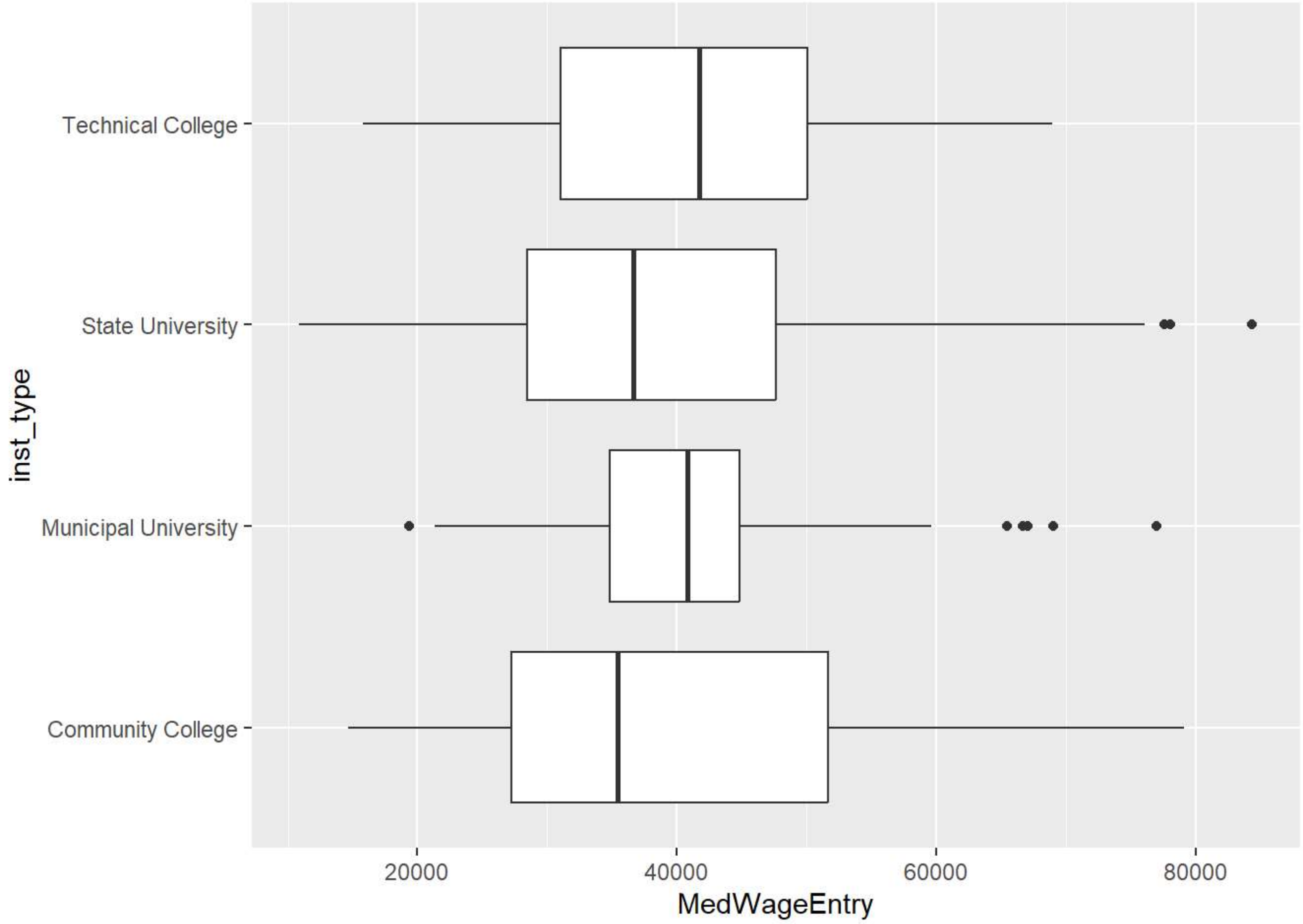
```
#find highest paying degrees
courses %>%
  group_by(DegreeTitle) %>%
  summarize(meanWage = mean(MedWage5yr, na.rm = TRUE)) %>%
  arrange(desc(meanWage))
```

```
## # A tibble: 434 × 2
##    DegreeTitle                        meanWage
##    <chr>                                 <dbl>
##  1 BUSINESS ANALYTICS                  107783
##  2 ELECTRICAL & POWER TRANSMISSION     103802
##  3 ARCHITECTURAL ENGINEERING           103643
##  4 ELECTRONICS ENGINEERING TECHNOLOGY  102894
##  5 COMPUTER ENGINEERING                101775.
##  6 APPLIED COMPUTING                   101604
##  7 ELECTRIC POWER AND DISTRIBUTION     101415
##  8 ELECTRICAL ENGINEERING               99578
##  9 COMPUTER SCIENCE                     95220.
## 10 PETROLEUM ENGINEERING                94728
## # ℹ 424 more rows
```

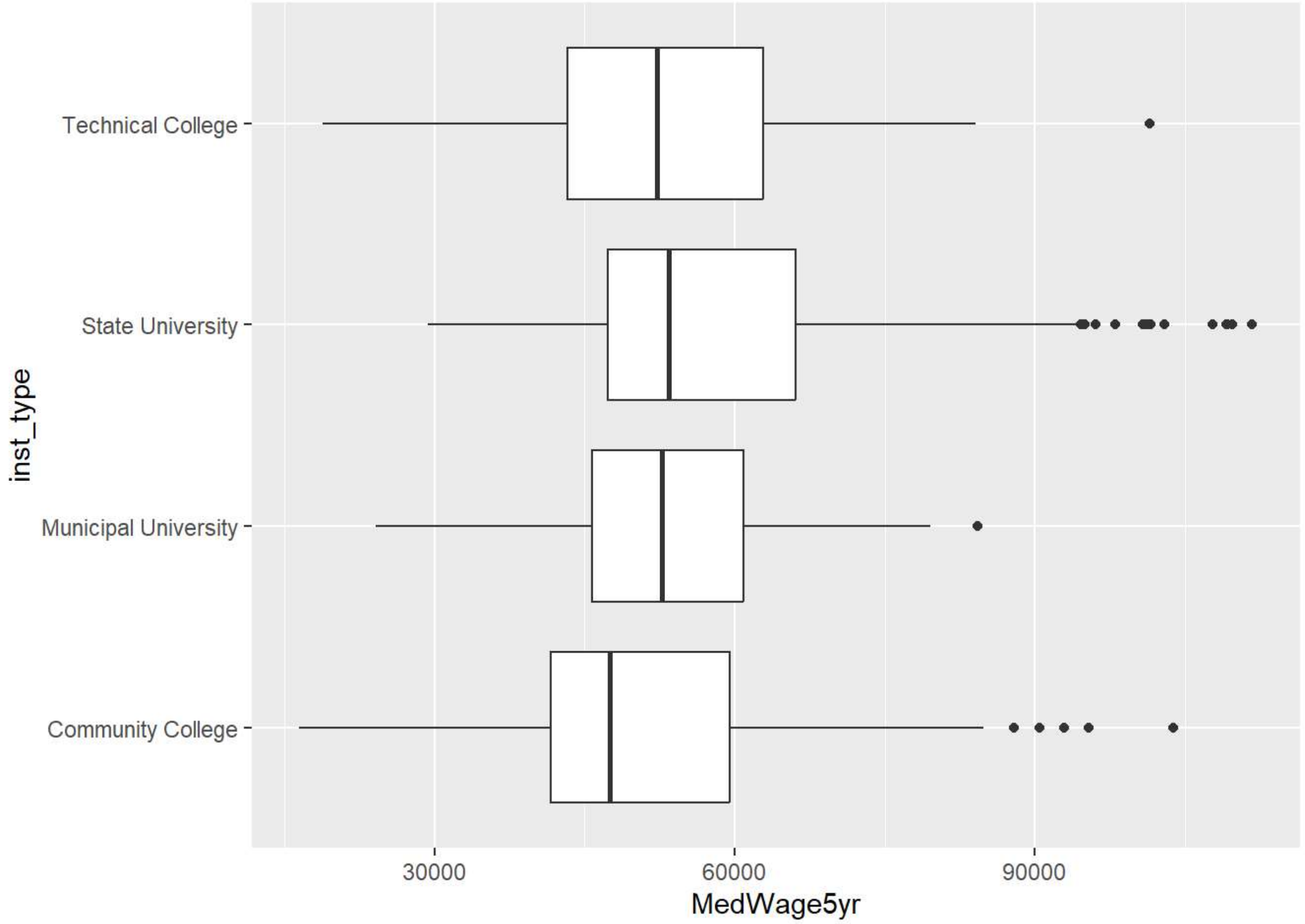
```
#find which degrees/programs have highest wage vs cost
courses %>%
  group_by(DegreeTitle) %>%
  summarize(meanValue = mean(Med5YrValue, na.rm = TRUE), meanWage = mean(MedWage5yr, na.rm = TRU
E), n = n()) %>%
  filter(n > 4) %>%
  arrange(desc(meanValue))
```

```
## # A tibble: 39 × 4
##    DegreeTitle                                         meanValue meanWage     n
##    <chr>                                                   <dbl>    <dbl> <int>
##  1 COSMETOLOGY                                              6.07   38076      5
##  2 LIBERAL ARTS AND SCIENCES, GENERAL STUDIES AND HUMA…     5.50   43052.    10
##  3 WELDING TECHNOLOGY                                       5.40   51550      7
##  4 AA, AS, AGS DEGREES (TRANSFER DEGREES)                   5.10   43240.    17
##  5 AUTOMOTIVE TECHNOLOGY                                    4.66   56143.    13
##  6 ELECTRICAL TECHNOLOGY                                    4.36   53807.     6
##  7 DIESEL TECHNOLOGY                                        4.24   54065.     5
##  8 PHYSICAL THERAPIST ASSISTANT                             4.12   49155      5
##  9 LIBERAL STUDIES                                          4.05   41600.     9
## 10 SURGICAL TECHNOLOGY                                      3.93   48051.     5
## # ℹ 29 more rows
```

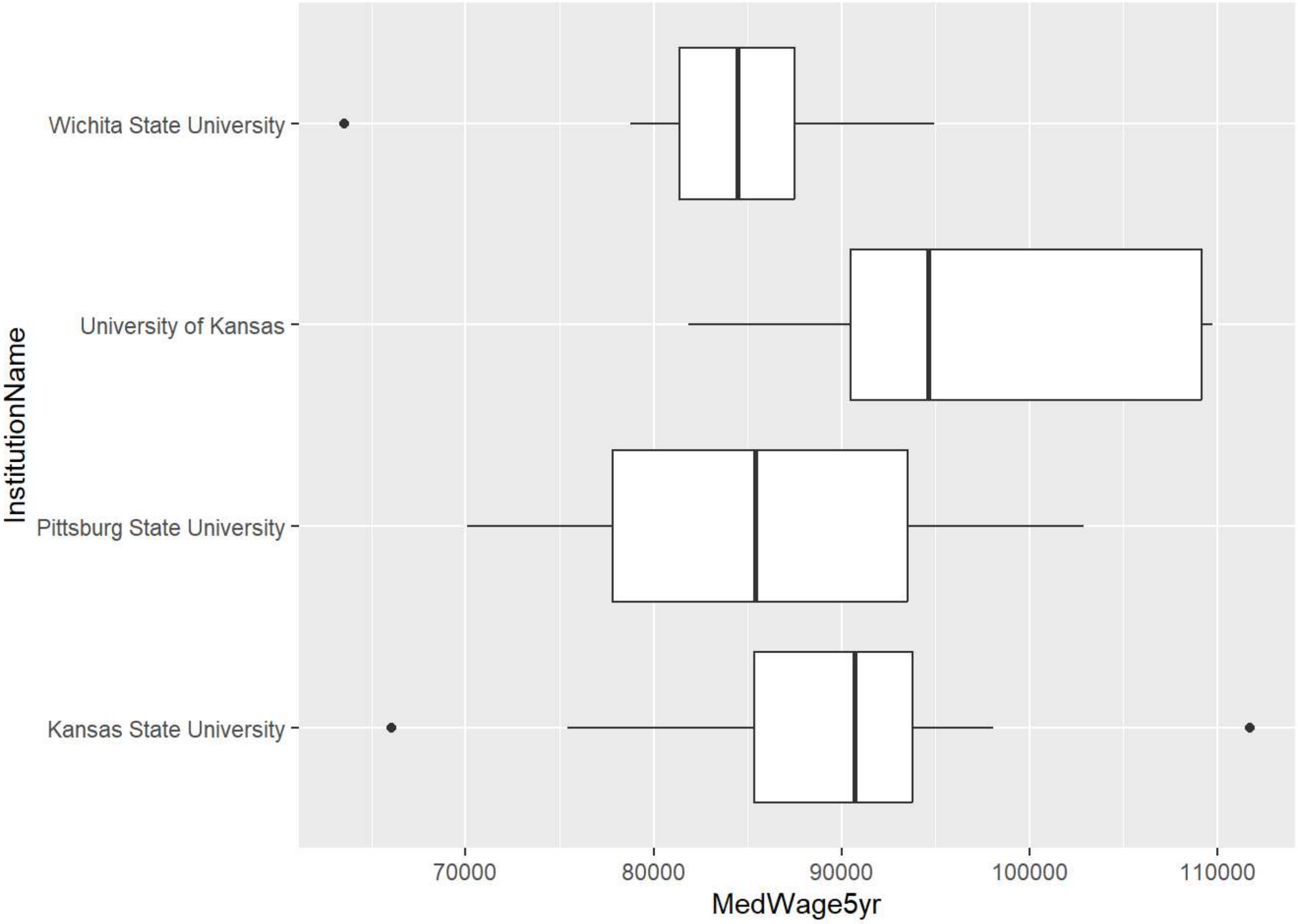
```
#compare wages by inst type
courses %>%
  left_join(institutions, by = "inst_id") %>%
  group_by(inst_type) %>%
  filter(!is.na(inst_type)) %>%
  ggplot() +
  geom_boxplot(mapping = aes(x = inst_type, y = MedWageEntry)) +
  coord_flip()
```



```
courses %>%
  left_join(institutions, by = "inst_id") %>%
  group_by(inst_type) %>%
  filter(!is.na(inst_type)) %>%
  ggplot() +
  geom_boxplot(mapping = aes(x = inst_type, y = MedWage5yr)) +
  coord_flip()
```



```
#compare only engineering degrees
courses %>%
  left_join(institutions, by = "inst_id") %>%
  filter(str_detect(DegreeTitle, "ENGINEERING")) %>%
  filter(inst_type %in% c("State University", "Municipal University")) %>%
  ggplot() +
  geom_boxplot(mapping = aes(x = InstitutionName, y = MedWage5yr)) +
  coord_flip()
```


Wichita State University
University of Kansas
Pittsburg State University
Kansas State University
InstitutionName
70000
80000
90000
100000
110000
MedWage5yr