ggplot Examples

Evan Bowman

22 March, 2023

College Scorecard

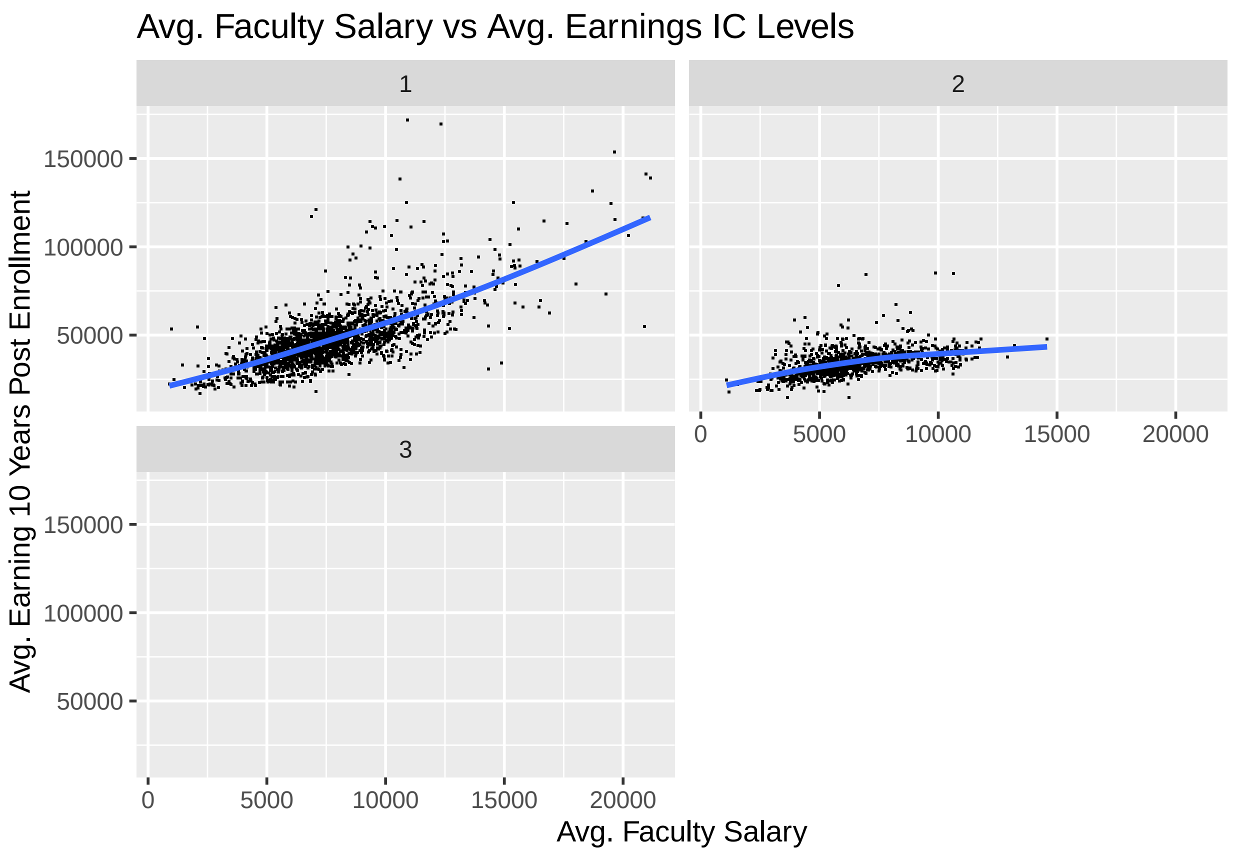
The data folder contains "college_scorecard_extract_2022-04-20.csv", a *subset* of the data in the College Scorecard database as of April 04. 2022. The data contain information on college cohorts in the United States. The data dictionary is in the `data` folder. The variables include:

- `UNITID` and `OPEID`: Identifiers for the colleges.
- `INSTNM`: Institution name
- `ADM_RATE`: The Admission Rate.
- `SAT_AVE`: Average SAT equivalent score of students admitted.
- `UGDS`: Enrollment of undergraduate certificate/degree-seeking students
- `COSTT4_A`: Average cost of attendance (academic year institutions)
- `AVGFACSAL`: Average faculty salary
- `GRAD_DEBT_MDN`: The median debt for students who have completed
- `AGE_ENTRY`: Average age of entry
- `ICLEVEL`: Level of institution (1 = 4-year, 2 = 2-year, 3 = less than 2-year).
- `MN_EARN_WNE_P10`: Mean earnings of students working and not enrolled 10 years after entry.
- `MD_EARN_WNE_P10`: Median earnings of students working and not enrolled 10 years after entry.
- `FEMALE`: Share of female students
- `PCT_WHITE`: Percent of the population from students' zip codes that is White, via Census data

4. How is average faculty salary associated with the average earnings of students ten years after initial enrollment?

- Use {ggplot2} to create an appropriate plot to assess for a relationship (with `AVGFACSAL` as the explanatory X variable), while using a {ggplot2} function argument to reduce over-plotting and adding the default smoother with `se = FALSE` and faceting on `ICLEVEL`.

```r
ggplot(scorecard, aes(AVGFACSAL, MN_EARN_WNE_P10))+
  geom_point(shape = ".")+
  geom_smooth(se = F)+
  facet_wrap(~ ICLEVEL, nrow = 2)+
  ggtitle("Avg. Faculty Salary vs Avg. Earnings IC Levels")+
  xlab("Avg. Faculty Salary")+
  ylab("Avg. Earning 10 Years Post Enrollment")
```



2. Interpret the plots about the potential relationship.

   In IC Level 1 Schools, there seems to be a clear curvilinear trend of average earnings increasing as the faculty salary increases. However, for IC Level 2 schools, there is a brief increase in average earnings but seems to flatten out very quickly. There is more range and variablility to the average earnings of IC Level 1 schools, whereas the average earnings for IC Level 2 schools seems to be clustered at lower earning levels. Furthermore, there is larger spread and varaiability with the average faculty salary in IC Level 1 schools when compared to the IC Level 2 schools.

3. Why is there no `ICLEVEL` 3 plot or if there is a plot why is there no data in the `ICLEVEL` 3 plot?

   As seen by the modified filter of the scorecard set below, the plot for IC Level 3 schools returns no observations because these schools did not report the average faculty salary.

```
scorecard%>%
  filter(ICLEVEL == 3) %>%
  summarise(Reported_Sal = sum(!is.na(AVGFACSAL)))
```

```
## # A tibble: 1 x 1
##   Reported_Sal
##          <int>
## 1            0
```

- Use `lm()` to run a linear model of the relationship **for only those schools with ICLEVEL 1** and save the results.

```
df1 <- scorecard %>%
  filter(ICLEVEL == 1)

lm1 <- lm(df1$MN_EARN_WNE_P10 ~ df1$AVGFACSAL)
summary(lm1)
```

```
##
## Call:
## lm(formula = df1$MN_EARN_WNE_P10 ~ df1$AVGFACSAL)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50732  -6461  -1054   4924 110064
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.379e+04  7.870e+02   17.53   <2e-16 ***
## df1$AVGFACSAL 4.394e+00  9.614e-02   45.70   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 11360 on 2050 degrees of freedom
##   (681 observations deleted due to missingness)
## Multiple R-squared:  0.5046, Adjusted R-squared:  0.5044
## F-statistic:  2088 on 1 and 2050 DF,  p-value: < 2.2e-16
```

1. Interpret the results of the model based on the p value and r-squared value.

   With a p-value close to zero, we reject the null hypothesis that there is no statistically significant relationship between average faculty salary and average earnings ten years following enrollment. With an r-squared value of .5044, only about 50 percent of the variation for average earnings is explained by the average faculty salary variable.

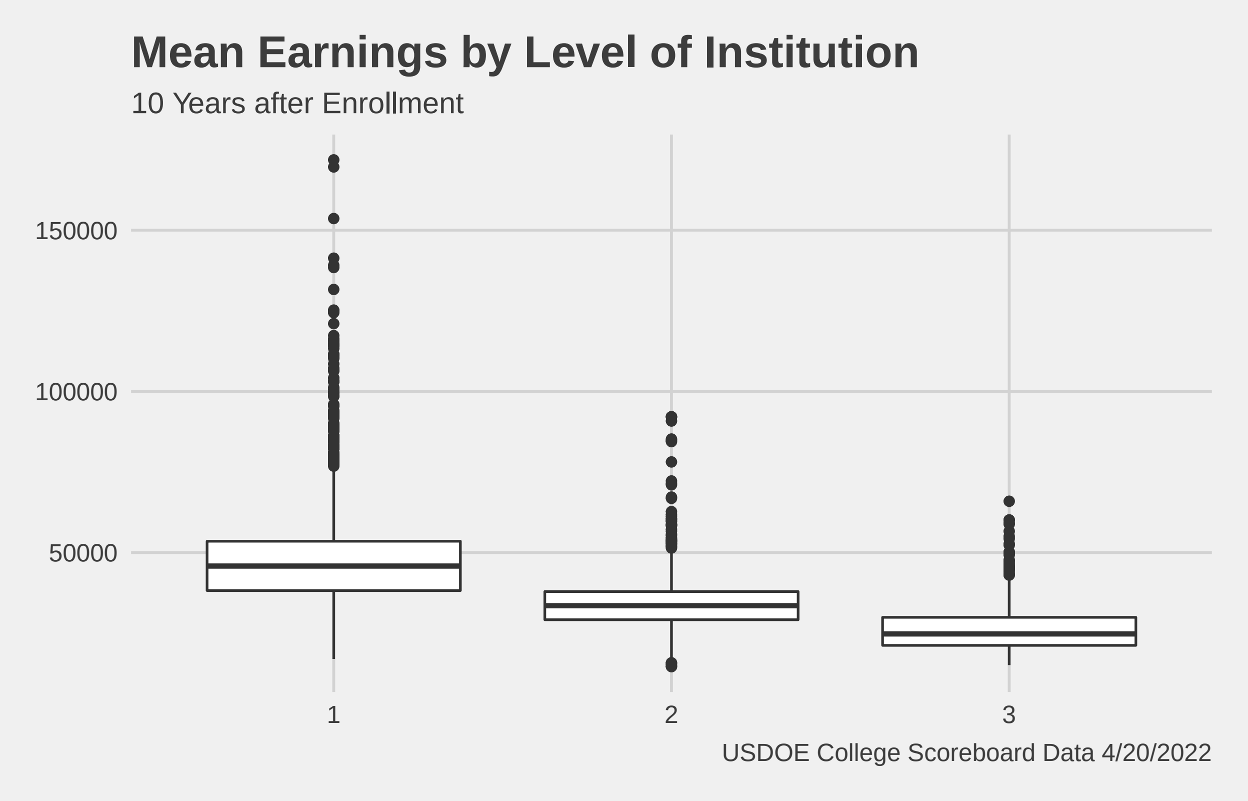
2. Given the r-squared value, what might you recommend to try to better predict average earnings of students ten years after initial enrollment at ICLEVEL 1 schools?

   I would first look at metrics that indicate the rigour (and thus reputation) of the school. Some of these variables would be admissions rate (ADM_RATE) and SAT average (SAT_AVE).

3. Does the level of the institution seem to be associated with the mean earnings of students ten years after enrollment?

- Make a plot

```
ggplot(scorecard)+
  geom_boxplot(aes(x = as.factor(ICLEVEL), y=MN_EARN_WNE_P10))+
  ggtitle("Mean Earnings by Level of Institution", subtitle = "10 Years after Enrollment")+
  labs(caption = "USDOE College Scoreboard Data 4/20/2022")+
  ggthemes::theme_fivethirtyeight()
```



2. Interpret the plot.

   The median of the average earnings for IC Level 1 schools are greater than both IC Level 2 and 3 schools. IC Level 2 schools also has higher median average earnings than IC Level 3 schools. IC Level 1 schools have higher variability in the outliers of average earnings (~ 100k) compared to IC Level 2 (~ 30k) and IC Level 3 (~30k).

- Use `aov()` to test if all of the IC levels have the same true mean of *logged* earnings of students ten years after enrollment.

```
anova <- aov(log(MN_EARN_WNE_P10) ~ ICLEVEL, data = scorecard)
```

1. Why would we look at the log of mean earnings instead of the un-logged values?

From the boxplot, there are a large number of outliers. The log transformation decreases the number of outliers by condensing the variability and spread of the data.

2. Use `broom::tidy()` to show the results.

```
broom::tidy(anova)
```

```
## # A tibble: 2 x 6
##   term         df sumsq   meansq statistic p.value
##   <chr>     <dbl> <dbl>    <dbl>     <dbl>   <dbl>
## 1 ICLEVEL       1  215. 215.         2827.       0
## 2 Residuals  4340  330.   0.0760        NA      NA
```

3. Interpret the results With a *p*-value of 0, we can confidently reject the null hypothesis that the mean average of earnings are equal across all IC Levels. There is statistically significant evidence that suggests at least one of the IC Level's mean average earnings are not equal to the others.

4. Create two plots with the attributes below

    1. Plot the median earnings 10 years after enrollment for level 1 institutions as the Y axis against `PCT_WHITE`
    2. Plot the median earnings 10 years after enrollment for level 1 institutions as the Y axis against `FEMALE`

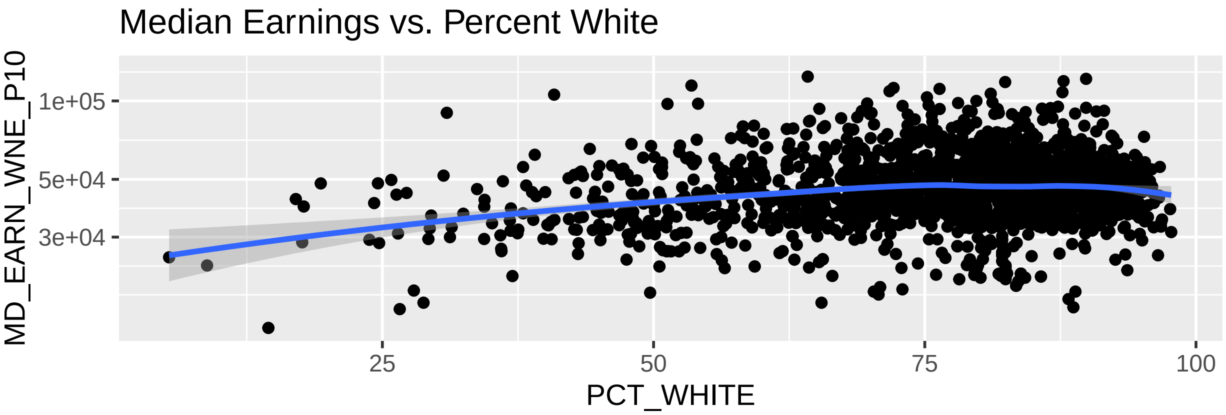
        - Use a log scale as appropriate.
        - Add a loess smoother.

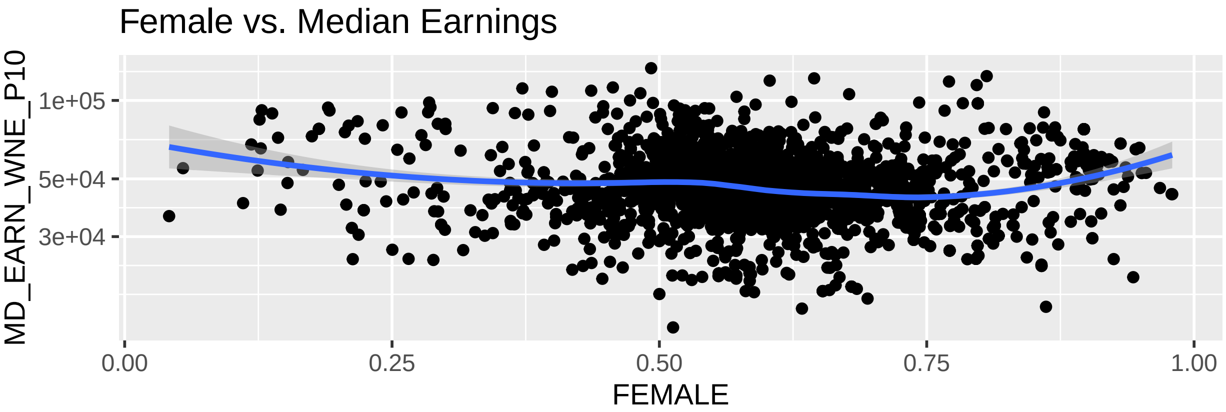
```r
# 6.1 Plot
df2 <- scorecard%>%
  filter(ICLEVEL == 1)

p_white <- ggplot(df2,aes(x =PCT_WHITE, y = MD_EARN_WNE_P10))+
  geom_point()+
  scale_y_log10() +
  geom_smooth(method = "loess")+
  ggtitle("Median Earnings vs. Percent White")

# 6.2 Plot
p_female <- ggplot(df2, aes(FEMALE, MD_EARN_WNE_P10))+
  geom_point()+
  scale_y_log10() +
  geom_smooth(method = "loess") +
  ggtitle("Female vs. Median Earnings")

gridExtra::grid.arrange(p_white, p_female)
```





3. Describe and interpret the non-linear relationship in each of the plots.

   For the plot analyziing median earnings against the percentaege of the white studetn population, there is an increase of median salary as the percentage of white students increases. This trend peaks at around 85 percent white student population and then begins to decrease at a larger magnitude than the previous increase. This shows that the majority of schools with higher median earnings are those that have a majority white presence on campus. This can fall into the conversation of funding for schools that serve more minority students which receive less funding and oppurunity, as well as the conversation regarding oppurtunity in public grade school education for predominately minority neighborhoods.

For the plot analyzing median earnings against the proportion of female students, there is a small curvilinear trend from 0 to around .52 that ultimately starts and ends at roughly the same log value of median earnings. As the proportion of females increases from .5, there is a strong negative relationship between proportion female and median earnings. This shows that the best schools for median earnings are those that have fairly equal representation between males and females in the student body.

4. What are the ethical implications of the definition of PCT_White?

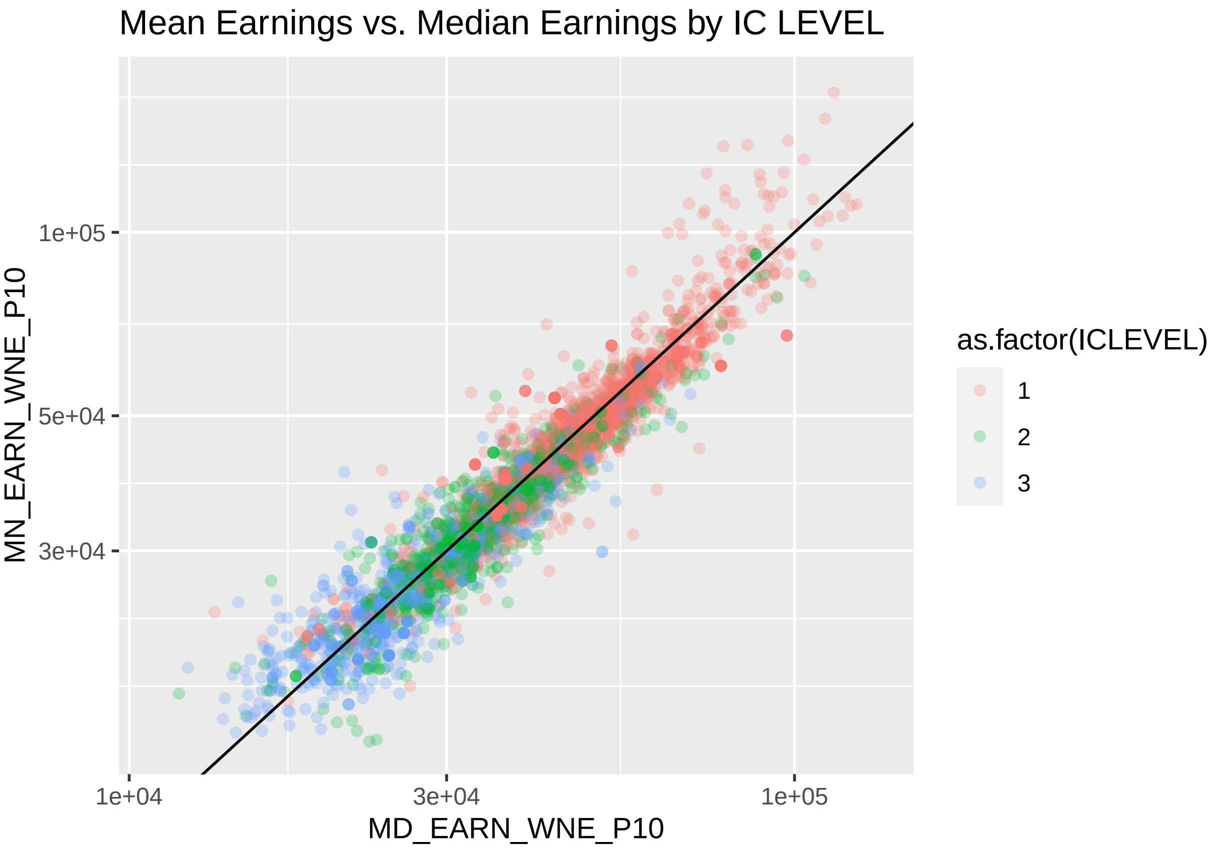
   There are several ethical implications of the definition of white. A major one is the analysis of the variable without taking into account any other variables. From just the first plot, there is a possible interpretation of median earnings increasing as the institution increaseness in whiteness.

5. Create a single scatter plot of the mean earnings 10 years after enrollment (Y axis) compared to the median earnings 10 years after enrollment (X axis) with the following attributes:

   1. Use log scales for both axes.

2. Add a 45 degree abline.
3. Use color to differentiate the level of institutions (as a factor) and
4. Attempt to reduce effects of over-plotting.
5. Interpret the plot and the relationship between the two variables **considering the abline**.

```
ggplot(scorecard, aes(MD_EARN_WNE_P10, MN_EARN_WNE_P10))+
  geom_point(aes(color = as.factor(ICLEVEL)), alpha = .25)+
  geom_abline(intercept = 0, slope = 1) +
  scale_x_log10() +
  scale_y_log10() +
  ggtitle("Mean Earnings vs. Median Earnings by IC LEVEL")
```



The plot shows the correlation between mean and median earnings for each IC Level and the discrepancies with each level given their main cluster on the plot. The abline also highlights how some institutions' mean earnings are inflated by a few graduates. Clusters around the line represent similar mean and median salaries across all graduates. Any straying from the line indicates the presence of specific outliers that are pushing the observations (schools) away from the balance between similar mean and median earnings.

8. Rankings Based on Return on Investment ROI

1. Calculate a measure of ROI as the ratio of median earnings 10 years after enrollment to median graduation debt for those schools that have both data elements - call it `ROI`.

```
scorecard_ROI <- scorecard%>%
  mutate(ROI = MD_EARN_WNE_P10 / GRAD_DEBT_MDN)
```

2. Use a {dplyr} function to compute a ranking of `ICLEVEL` 1 universities based on `ROI`. - Break any ties using the minimum value such that if two schools are tied for 4th, they are each ranked 4 and the next school is 6.

```
scorecard_ROI_rank <-scorecard_ROI%>%
  filter(ICLEVEL == 1)%>%
  mutate(Rank = min_rank(desc(ROI)))%>%
  arrange(Rank)
```

3. Identify the top 5 best (highest ROI should be rank value 1) and the bottom 5 worst (smallest ROI should have largest rank number). Show only the rank, name, ROI, cost to attend, median debt at graduation, and median earnings 10 years after enrollment.

```
# 5 Best
scorecard_ROI_rank%>%
  select(Rank, INSTNM ,ROI, COSTT4_A, GRAD_DEBT_MDN, MD_EARN_WNE_P10)%>%
  head(5)
```

```
## # A tibble: 5 x 6
##    Rank INSTNM                                  ROI COSTT4_A GRAD_~1 MD_EA~2
##   <int> <chr>                                 <dbl>    <dbl>   <dbl>   <dbl>
## 1     1 Laredo College                        14.1     11250    2334   33019
## 2     2 Berea College                         10.0     54866    3700   37154
## 3     3 SUNY Downstate Health Sciences University  9.16       NA   12500  114551
## 4     4 Princeton University                   9.16    74150   10450   95689
## 5     5 Franklin W Olin College of Engineering  8.39    74286   15846  132969
## # ... with abbreviated variable names 1: GRAD_DEBT_MDN, 2: MD_EARN_WNE_P10
```

```
# 5 Worst
scorecard_ROI_rank%>%
  arrange(ROI)%>%
  select(Rank, INSTNM ,ROI, COSTT4_A, GRAD_DEBT_MDN, MD_EARN_WNE_P10)%>%
  head(5)
```

```
## # A tibble: 5 x 6
##    Rank INSTNM                      ROI COSTT4_A GRAD_DEBT_MDN MD_EARN_WNE_P10
##   <int> <chr>                     <dbl>    <dbl>         <dbl>           <dbl>
## 1  2290 American Baptist College  0.538    24972         43000           23135
## 2  2289 Martin University         0.657    23393         41604           27347
## 3  2288 The North Coast College   0.714    28267         48148           34388
## 4  2287 Nightingale College       0.768       NA         21000           16130
## 5  2285 Eagle Gate College-Murray 0.792    22097         41639           32969
```

4. What is American University's rank and ROI? Show only the rank, name, ROI, cost to attend, median debt at graduation and median earnings 10 years after enrollment.

```
scorecard_ROI_rank%>%
  filter(INSTNM == "American University")%>%
  select(Rank, INSTNM, ROI, COSTT4_A, GRAD_DEBT_MDN, MD_EARN_WNE_P10)
```

```
## # A tibble: 1 x 6
##    Rank INSTNM                ROI COSTT4_A GRAD_DEBT_MDN MD_EARN_WNE_P10
##   <int> <chr>               <dbl>    <dbl>         <dbl>           <dbl>
## 1   455 American University  3.09    66416         23250           71933
```

American's ROI ranking is 455.

2. What is AU's new ranking and ROI if the mean earnings are used?

```
scorecard_ROI_mean <- scorecard%>%
  mutate(ROI = MN_EARN_WNE_P10 / GRAD_DEBT_MDN)

scorecard_ROI_mean_rank <-scorecard_ROI_mean%>%
  filter(ICLEVEL == 1)%>%
  mutate(Rank = min_rank(desc(ROI)))%>%
  arrange(Rank)

scorecard_ROI_mean_rank%>%
  filter(INSTNM == "American University")%>%
  select(Rank, INSTNM, ROI, COSTT4_A, GRAD_DEBT_MDN, MN_EARN_WNE_P10)
```

```
## # A tibble: 1 x 6
##    Rank INSTNM                ROI COSTT4_A GRAD_DEBT_MDN MN_EARN_WNE_P10
##   <int> <chr>               <dbl>    <dbl>         <dbl>           <dbl>
## 1   432 American University  2.91    66416         23250           67700
```

Using the mean earnings ten years post enrollment, AU's ranking increased from 455 to 432.

World Bank Data

The World Bank provides loans to countries with the goal of reducing poverty. The data frames in the data folder were taken from the public data repositories of the World Bank.

- `country_2021.csv`: Contains information on the countries in the data set. Also includes totals for the regions (sets of countries) and the world. The variables are:
  - `Country_Code`: A three-letter code for the country. Note not all rows are countries; There are rows for the regions and for the world.
  - `Region`: The region of the country. **Take note: Region is `NA` for the individual regions and for the world.**
  - `IncomeGroup`: Either `"High income"`, `"Upper middle income"`, `"Lower middle income"`, or `"Low income"`. Take note: Region is `NA` for the individual regions and for the world.
  - `TableName`: The full name of the country or region.
  - `Special Notes`: Notes about the country or region
- `fertility_rate_2020.csv`: Contains the fertility rate information for each country, each region, and the world for each year.
  - For the variables `1960` to `2020`, the values in the cells represent the fertility rate in total births per woman for that year.
  - Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.
- `life_expectancy_2020.csv`: Contains the life expectancy information for each country, each region, and the world for each year.
  - For the variables `1960` to `2020`, the values in the cells represent life expectancy at birth in years for the given year.

    - Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.

- `population_2021.csv`: Contains the population information for each country, each region, and the world for the year.
    - For the variables `1960` to `2021`, the values in the cells represent the total population in number of people for the given year.
    - Total population is based on the *de facto* definition of population, which counts all residents regardless of legal status or citizenship. The values shown are mid-year estimates.

2. These data are messy. The observational units in `fert`, `life`, and `pop` are locations in space-time (e.g. Aruba in 2020). Recall that tidy data should have one observational unit per row. Using only two function calls for each data frame, tidy each data frame to have only four variables by:
    1. Removing the `Indicator Name` and `Indicator Code` columns.
    2. Using a {tidyr} function to tidy the tibble, and, by using an argument, ensure the variable for `year` is a numeric.
    3. Save the tidy tibble to a new name.

```r
# Fertility Data
df <- fertility %>%
  select(-3 , -4)

fert_tidy <- df %>%
  pivot_longer(
    cols = 3:64,
    names_to = c("Year"),
    values_to = "Fert_Rate",
    names_transform = list(Year = as.numeric))

# Life Expectancy Data
df1 <- life %>%
  select(-3, -4)

life_tidy <- df1%>%
  pivot_longer(
    cols = 3:64,
    names_to = "Year",
    values_to = "Life_Expect",
    names_transform = list(Year = as.numeric))

# Population Data
df2 <- population %>%
  select(-3, -4)

pop_tidy <- df2%>%
  pivot_longer(
    cols = 3:64,
    names_to = "Year",
    values_to = "Population",
    names_transform = list(Year = as.numeric))
```

3. Combine tibbles.

1. Using a {dplyr} function, *join* the three tidy tibbles into one tibble, one after the other, and then use a {dplyr} function to join the data from the country tibble while removing any rows that have no data in country. The new tibble should have 16430 observations (rows) for 10 variables.

```
fert_tidy %>%
  right_join(life_tidy, by = NULL) %>%
  right_join(pop_tidy, by = NULL) %>%
  right_join(country, by = NULL) %>%
  tibble()
```

```
## # A tibble: 16,430 x 10
##    Countr~1 Count~2  Year Fert_~3 Life_~4 Popul~5 Region Incom~6 Table~7 Speci~8
##    <chr>    <chr>   <dbl>   <dbl>   <dbl>   <dbl> <chr>  <chr>   <chr>   <chr>
##  1 Aruba    ABW      1960    4.82    65.7   54208 Latin~ High i~ Aruba   <NA>
##  2 Aruba    ABW      1961    4.66    66.1   55434 Latin~ High i~ Aruba   <NA>
##  3 Aruba    ABW      1962    4.47    66.4   56234 Latin~ High i~ Aruba   <NA>
##  4 Aruba    ABW      1963    4.27    66.8   56699 Latin~ High i~ Aruba   <NA>
##  5 Aruba    ABW      1964    4.06    67.1   57029 Latin~ High i~ Aruba   <NA>
##  6 Aruba    ABW      1965    3.84    67.4   57357 Latin~ High i~ Aruba   <NA>
##  7 Aruba    ABW      1966    3.62    67.8   57702 Latin~ High i~ Aruba   <NA>
##  8 Aruba    ABW      1967    3.42    68.1   58044 Latin~ High i~ Aruba   <NA>
##  9 Aruba    ABW      1968    3.23    68.4   58377 Latin~ High i~ Aruba   <NA>
## 10 Aruba    ABW      1969    3.05    68.8   58734 Latin~ High i~ Aruba   <NA>
## # ... with 16,420 more rows, and abbreviated variable names 1: ‘Country Name‘,
## #   2: ‘Country Code‘, 3: Fert_Rate, 4: Life_Expect, 5: Population,
## #   6: IncomeGroup, 7: TableName, 8: SpecialNotes
```

3. Identify the distinct values for the two columns of `Country Name` and `TableName` where they do not match each other. There should be nine. What do you notice about them?

```
differ <- subset(full_tidy, !(TableName %in% `Country Name`))
unique(differ$`Country Name`)
```

```
## [1] "Cote d’Ivoire"
## [2] "Curacao"
## [3] "Sao Tome and Principe"
## [4] "East Asia & Pacific (IDA & IBRD countries)"
## [5] "Europe & Central Asia (IDA & IBRD countries)"
## [6] "Latin America & the Caribbean (IDA & IBRD countries)"
## [7] "Middle East & North Africa (IDA & IBRD countries)"
## [8] "Sub-Saharan Africa (IDA & IBRD countries)"
## [9] "Turkiye"
```

```
unique(differ$TableName)
```

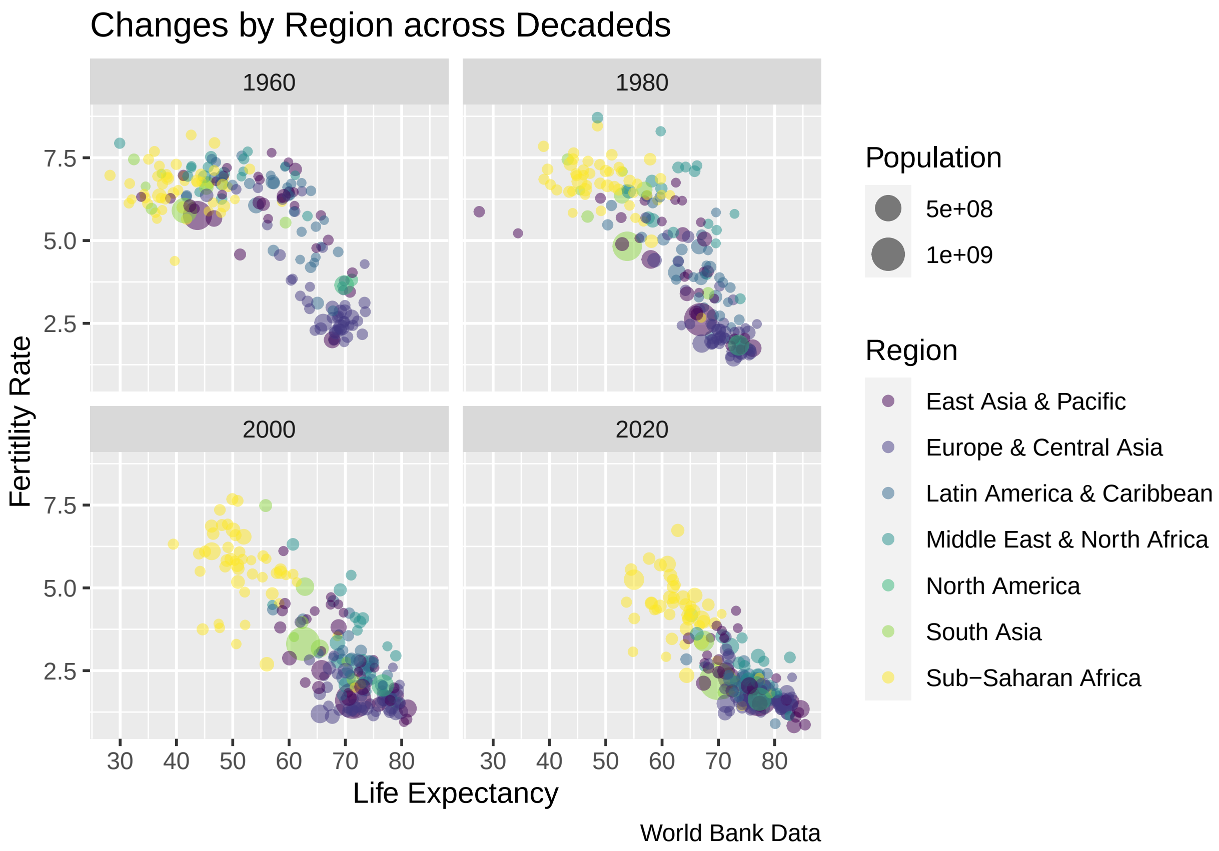
```
## [1] "Côte d’Ivoire"
## [2] "Curaçao"
## [3] "São Tomé and Principe"
## [4] "East Asia & Pacific (IDA & IBRD)"
## [5] "Europe & Central Asia (IDA & IBRD)"
## [6] "Latin America & Caribbean (IDA & IBRD)"
## [7] "Middle East & North Africa (IDA & IBRD)"
## [8] "Sub-Saharan Africa (IDA & IBRD)"
## [9] "Türkiye"
```

The nine distinct values are countries and regions that differ in spelling between the Country Name and Table Name columns. These are discrepancies in spelling (i.e. accents on letters for Table Name that are not in the Country Name titles) or parenthesis clauses that are used in Country Name and not Table Name.

4. Make a single scatterplot with the following attributes:
    1. Show fertility rate (Y) vs life expectancy (X),
    2. Color-code by region and indicate population by size.
    3. Include only the years 1960, 1980, 2000, and 2020.
    4. Facet by these years.
    5. **Interpret the plot in one sentence**.

```
target <- c(1960, 1980, 2000, 2020)
facet_year <- full_tidy%>%
  filter(Year %in% target)

facet_year%>%
  filter(Region != is.na(Region))%>%
ggplot(aes(Life_Expect, Fert_Rate, size = Population, color = Region))+
  geom_point(alpha = .5)+
  facet_wrap(~ as.factor(Year))+
  ggtitle("Changes by Region across Decadeds")+
  xlab("Life Expectancy")+
  ylab("Fertitlity Rate")+
  labs(caption = "World Bank Data")+
  scale_color_viridis_d()
```

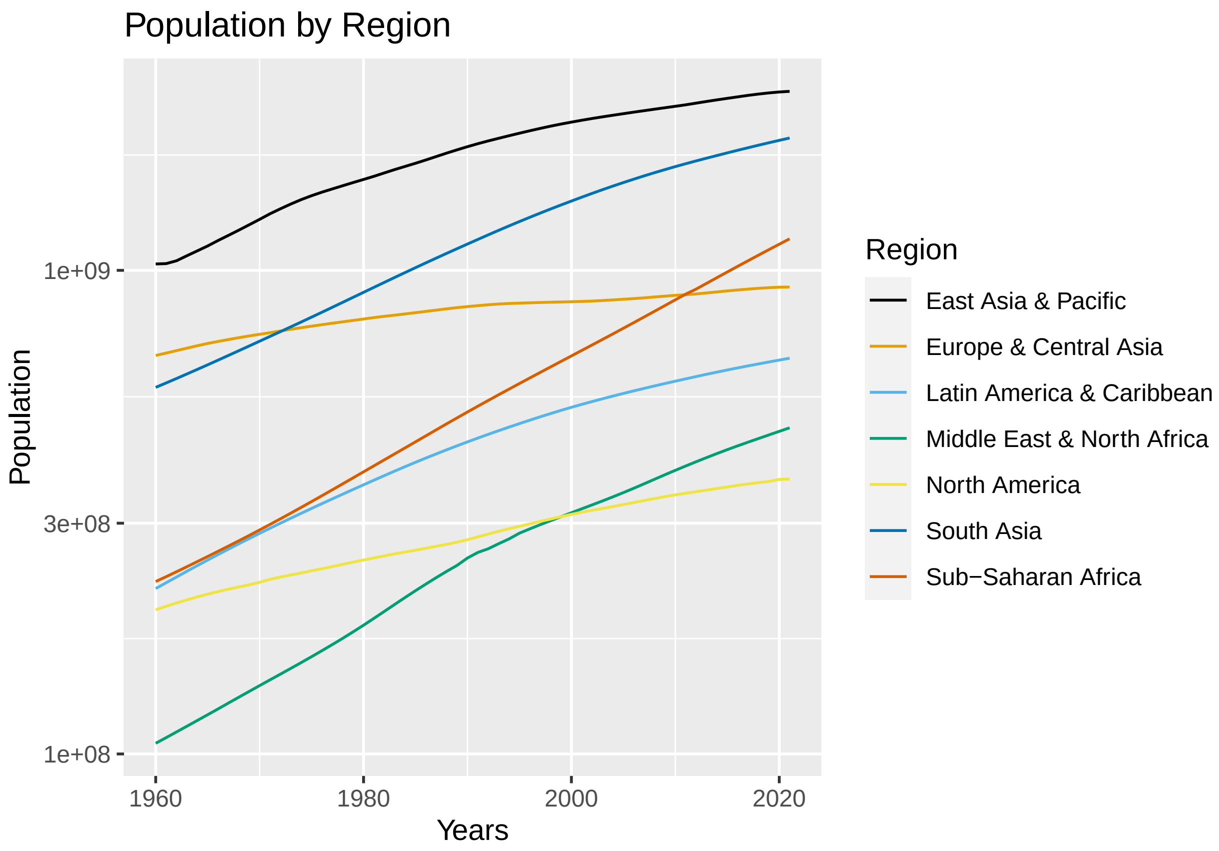


There is a clear negative association between life expectancy and fertility rate. There are also some interesting changes over the decades. In all four decades, fertility rate as consistently decreased while life expectancy has decreased. There was smaller discrepancies between fertility rate and life expectancy between Sub-Saharan Africa and the rest of the world in 1960. However, as time progressed, other regions of the world seemed to advance in life expectancy and decrease in fertility rate at a larger rate than the region. This highlights the lack of development in the region. The region that seems to have developed the most in the last sixty years in East Asia and the Pacific.

5. Regional Population
    1. Calculate the total population for each region for each year.
    2. Make a line plot of year (Y) versus total population (X), color-coding by region and using a *log scale* for Y.
    3. **Interpret the plot in one sentence to identify the two fastest growing regions**.

```
pop_year <- full_tidy%>% group_by(Region)%>%
  count(Year, wt = Population, sort = T)

pop_year%>%
  filter(Region != is.na(Region))%>%
  ggplot(aes(Year, n, color = Region))+
  geom_line()+
  ggtitle("Population by Region")+
  xlab("Years")+
  ylab("Population")+
  scale_y_log10()+
  ggthemes::scale_color_colorblind()
```

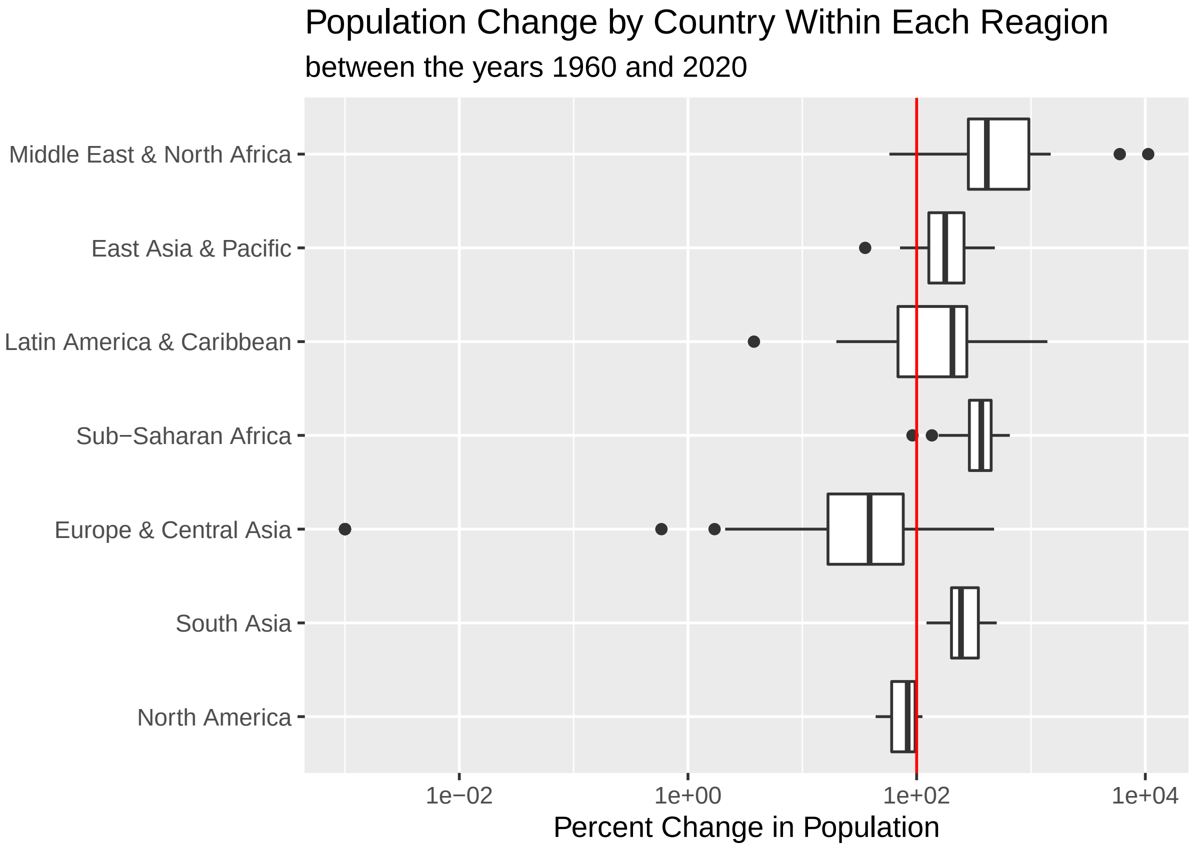


The two fastest growing regions are Sub-Saharan Africa and the Middle East and North Africa.

6. Make a box plot of the percentage population growth for the countries in each region in the sixty year period between the two years 1960 and 2020 with the following attributes.
    1. Use code to automatically order the Regions on the y-axis in **increasing** order of total 1960 population.
    2. For any countries with negative growth set to .001.
    3. Add a red line at 100%
    4. **Interpret the plot in one sentence to compare the percentage growth with the previous plot.**

```
target1 <- c(1960, 2020)
growth_prop <- full_tidy%>% group_by(`Country Name`)%>%
  filter(Region != is.na(Region))%>%
  filter(Year %in% target1)%>%
  mutate(diff_year = (Year - lead(Year)) * -1,
         diff_pop = (Population - lead(Population)) * -1,
         percent_growth = (diff_pop - diff_year) / Population * 100)
growth_prop$percent_growth[growth_prop$percent_growth < 0] <- .001

growth_prop$Region = as.factor(growth_prop$Region)
growth_prop%>% group_by(Region)%>%
  filter(Region != is.na(Region), Year == 1960)%>%
  ggplot(aes(fct_reorder(Region, Population, .fun = median, .desc = T), percent_growth))+
  geom_boxplot()+
  coord_flip()+
  scale_y_log10()+
  geom_hline(yintercept = 100, color = "red")+
  ggtitle("Population Change by Country Within Each Reagion", subtitle = "between the years 1960 and 20
  xlab("")+
  ylab("Percent Change in Population")
```



The boxplot confirms the Middle East & North Africa and Sub-Saharan Africa had the largest growth rate over the 60 year period.