Finance and Insurance Summary Statistics

1. Source: <https://www.census.gov/data/tables/2017/econ/economic-census/naics-sector-52.html>
2. Download data from link above.
3. This downloads as a .DAT file and will need to be opened in a text editor and saved as a .txt file.
4. Open Excel and ‘get data’ from the text file you just saved. The delimiter is a pipe (’|’) and Excel should recognize this and add this to the delimiter drop down.
5. The data types of the columns should be recognized automatically and assigned as such.
6. Remove the columns with the following names:
   1. "GEOTYPE"
   2. "ST"
   3. "COUNTY"
   4. "PLACE"
   5. "CONSCITY"
   6. "CSA"
   7. "MSA"
   8. "MD"
   9. "GEOCOMP"
   10. "GEO\_ID"
   11. "GEO\_ID\_F"
   12. "NAICS2017\_F"
   13. "FIRM\_F"
   14. "ESTAB\_F"
   15. "RCPTOT\_F"
   16. "PAYANN\_F"
   17. "PAYQTR1\_F"
   18. "EMP\_F"
   19. "RCPTOT\_IMP\_F"
   20. "PAYANN\_IMP\_F"
   21. "EMP\_IMP\_F"
7. Highlight the GEO\_TTL column and format the text so that each word is capitalized.
8. Add a column named AVG SALARY(PAYANN/EMP) -- Round to 2 decimals
   1. Number.Round([PAYANN]/[EMP], 2)
9. Filter out NaN and infinity in the newly created column
10. Save and Load
11. Advanced view of Power Query:
    1. let
    2. Source = Csv.Document(File.Contents("C:\Users\JRO20\OneDrive\Desktop\Dev10\Data Cohort\Capstone\Finanace\_Insurance\_Summary\_Stats.txt"),[Delimiter="|", Columns=35, Encoding=1252, QuoteStyle=QuoteStyle.None]),
    3. #"Promoted Headers" = Table.PromoteHeaders(Source, [PromoteAllScalars=true]),
    4. #"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"GEOTYPE", Int64.Type}, {"ST", Int64.Type}, {"COUNTY", Int64.Type}, {"PLACE", Int64.Type}, {"CONSCITY", type text}, {"CSA", Int64.Type}, {"MSA", Int64.Type}, {"MD", Int64.Type}, {"GEOCOMP", Int64.Type}, {"GEO\_ID", type text}, {"GEO\_TTL", type text}, {"GEO\_ID\_F", type text}, {"SECTOR", Int64.Type}, {"NAICS2017", Int64.Type}, {"NAICS2017\_TTL", type text}, {"NAICS2017\_F", type text}, {"YEAR", Int64.Type}, {"FIRM", Int64.Type}, {"FIRM\_F", type text}, {"ESTAB", Int64.Type}, {"ESTAB\_F", type text}, {"RCPTOT", Int64.Type}, {"RCPTOT\_F", type text}, {"PAYANN", Int64.Type}, {"PAYANN\_F", type text}, {"PAYQTR1", Int64.Type}, {"PAYQTR1\_F", type text}, {"EMP", Int64.Type}, {"EMP\_F", type text}, {"RCPTOT\_IMP", Int64.Type}, {"RCPTOT\_IMP\_F", type text}, {"PAYANN\_IMP", Int64.Type}, {"PAYANN\_IMP\_F", type text}, {"EMP\_IMP", Int64.Type}, {"EMP\_IMP\_F", type text}}),
    5. #"Removed Columns" = Table.RemoveColumns(#"Changed Type",{"GEOTYPE", "ST", "COUNTY", "PLACE", "CONSCITY", "CSA", "MSA", "MD", "GEOCOMP", "GEO\_ID", "GEO\_ID\_F", "NAICS2017\_F", "FIRM\_F", "ESTAB\_F", "RCPTOT\_F", "PAYANN\_F", "PAYQTR1\_F", "EMP\_F", "RCPTOT\_IMP\_F", "PAYANN\_IMP\_F", "EMP\_IMP\_F"}),
    6. #"Capitalized Each Word" = Table.TransformColumns(#"Removed Columns",{{"GEO\_TTL", Text.Proper, type text}})
    7. in
    8. #"Capitalized Each Word"
12. Saved as csv file and put into a pandas data frame
13. Labeled which layer each row was a part of because a lot of our data was just a breakdown further into a previous row.
    1. df['layer'] = df.NAICS2017.apply(lambda x: 7 if x>=1000000 and x<=9999999 else 6 if x>=100000 and x<=999999 else 5 if x>=10000 and x<=99999 else 4 if x>=1000 and x<=9999 else 3 if x>=100 and x<=999 else 2 if x>=10 and x<=99 else 0)
14. Created a separate df with only layer 6 for ease of analysis
15. Data Dictionary is located in the GitHub

School Data

1. Source: <https://www.census.gov/data/tables/2017/econ/school-finances/secondary-education-finance.html>
2. Data Dictionary and ENROLL Column source: <https://www2.census.gov/programs-surveys/school-finances/tables/2017/secondary-education-finance/>
3. Download both Data sheets from the source.
4. Download the Data Dictionary from the source.
5. Open excel and start a query with the downloaded school data.
6. Change the type of STATE to text
7. Capitalize each word in NAME
8. Replace values in NAME ‘ School District’ to ‘,’ (include the space before ‘School’)
9. Open a new excel file and the data dictionary you downloaded.
10. Copy and paste the State names and their codes from the data dictionary into a blank excel file and save it.
11. Go back to your first excel file and start a new query from the State data you just saved.
12. Transform the State Code data type to text.
13. Merge the queries together on the STATE and State Code columns. We only need the State name from the State data.
14. Merge the NAME and State Name columns by putting a space as the separator. This will create a new column at the end of your dataframe.
15. Slide that new column back to the front right behind IDCENSUS.
16. Open a new query and query the elsec17t sheet.
17. Merge the queries on the IDCENSUS column.
18. Only keep the following columns from elsec17t:
    1. ENROLL
    2. TOTALREV
    3. TFEDREV
    4. TSTREV
    5. TLOCREV
    6. TOTALEXP
    7. TOTALCURSPND
    8. TCURINST
    9. TCURSSVC
    10. PPCSTOT
    11. PPSALWG
    12. PPITOTAL
    13. PPSTOTAL
19. Save and Load.
20. Advanced view of Power Query:
    1. let
    2. Source = Excel.CurrentWorkbook(){[Name="Table1"]}[Content],
    3. #"Changed Type" = Table.TransformColumnTypes(Source,{{"STATE", type text}, {"IDCENSUS", Int64.Type}, {"NAME", type text}, {"CONUM", Int64.Type}, {"CSA", type text}, {"CBSA", type text}, {"SCHLEV", Int64.Type}, {"NCESID", Int64.Type}, {"YRDATA", Int64.Type}, {"V33", Int64.Type}, {"TOTALREV", Int64.Type}, {"TFEDREV", Int64.Type}, {"C14", Int64.Type}, {"C15", Int64.Type}, {"C16", Int64.Type}, {"C17", Int64.Type}, {"C19", Int64.Type}, {"B11", Int64.Type}, {"C20", Int64.Type}, {"C25", Int64.Type}, {"C36", Int64.Type}, {"B10", Int64.Type}, {"B12", Int64.Type}, {"B13", Int64.Type}, {"TSTREV", Int64.Type}, {"C01", Int64.Type}, {"C04", Int64.Type}, {"C05", Int64.Type}, {"C06", Int64.Type}, {"C07", Int64.Type}, {"C08", Int64.Type}, {"C09", Int64.Type}, {"C10", Int64.Type}, {"C11", Int64.Type}, {"C12", Int64.Type}, {"C13", Int64.Type}, {"C24", Int64.Type}, {"C35", Int64.Type}, {"C38", Int64.Type}, {"C39", Int64.Type}, {"TLOCREV", Int64.Type}, {"T02", Int64.Type}, {"T06", Int64.Type}, {"T09", Int64.Type}, {"T15", Int64.Type}, {"T40", Int64.Type}, {"T99", Int64.Type}, {"D11", Int64.Type}, {"D23", Int64.Type}, {"A07", Int64.Type}, {"A08", Int64.Type}, {"A09", Int64.Type}, {"A11", Int64.Type}, {"A13", Int64.Type}, {"A15", Int64.Type}, {"A20", Int64.Type}, {"A40", Int64.Type}, {"U11", Int64.Type}, {"U22", Int64.Type}, {"U30", Int64.Type}, {"U50", Int64.Type}, {"U97", Int64.Type}, {"TOTALEXP", Int64.Type}, {"TCURELSC", Int64.Type}, {"TCURINST", Int64.Type}, {"E13", Int64.Type}, {"J13", Int64.Type}, {"J12", Int64.Type}, {"J14", Int64.Type}, {"V91", Int64.Type}, {"V92", Int64.Type}, {"TCURSSVC", Int64.Type}, {"E17", Int64.Type}, {"E07", Int64.Type}, {"E08", Int64.Type}, {"E09", Int64.Type}, {"V40", Int64.Type}, {"V45", Int64.Type}, {"V90", Int64.Type}, {"V85", Int64.Type}, {"J17", Int64.Type}, {"J07", Int64.Type}, {"J08", Int64.Type}, {"J09", Int64.Type}, {"J40", Int64.Type}, {"J45", Int64.Type}, {"J90", Int64.Type}, {"J11", Int64.Type}, {"J96", Int64.Type}, {"TCUROTH", Int64.Type}, {"E11", Int64.Type}, {"V60", Int64.Type}, {"V65", Int64.Type}, {"J10", Int64.Type}, {"J97", Int64.Type}, {"NONELSEC", Int64.Type}, {"V70", Int64.Type}, {"V75", Int64.Type}, {"V80", Int64.Type}, {"J98", Int64.Type}, {"TCAPOUT", Int64.Type}, {"F12", Int64.Type}, {"G15", Int64.Type}, {"K09", Int64.Type}, {"K10", Int64.Type}, {"K11", Int64.Type}, {"J99", Int64.Type}, {"L12", Int64.Type}, {"M12", Int64.Type}, {"Q11", Int64.Type}, {"I86", Int64.Type}, {"Z32", Int64.Type}, {"Z33", Int64.Type}, {"V11", Int64.Type}, {"V13", Int64.Type}, {"V15", Int64.Type}, {"V17", Int64.Type}, {"V21", Int64.Type}, {"V23", Int64.Type}, {"V37", Int64.Type}, {"V29", Int64.Type}, {"Z34", Int64.Type}, {"V10", Int64.Type}, {"V12", Int64.Type}, {"V14", Int64.Type}, {"V16", Int64.Type}, {"V18", Int64.Type}, {"V22", Int64.Type}, {"V24", Int64.Type}, {"V38", Int64.Type}, {"V30", Int64.Type}, {"V32", Int64.Type}, {"\_19H", Int64.Type}, {"\_21F", Int64.Type}, {"\_31F", Int64.Type}, {"\_41F", Int64.Type}, {"\_61V", Int64.Type}, {"\_66V", Int64.Type}, {"W01", Int64.Type}, {"W31", Int64.Type}, {"W61", Int64.Type}}),
    4. #"Capitalized Each Word" = Table.TransformColumns(#"Changed Type",{{"NAME", Text.Proper, type text}}),
    5. #"Replaced Value" = Table.ReplaceValue(#"Capitalized Each Word"," School District",",",Replacer.ReplaceText,{"NAME"}),
    6. #"Merged Queries" = Table.NestedJoin(#"Replaced Value", {"STATE"}, States, {"State Number"}, "Sheet1", JoinKind.LeftOuter),
    7. #"Expanded Sheet1" = Table.ExpandTableColumn(#"Merged Queries", "Sheet1", {"Sate Name"}, {"Sate Name"}),
    8. #"Merged Columns" = Table.CombineColumns(#"Expanded Sheet1",{"NAME", "Sate Name"},Combiner.CombineTextByDelimiter(" ", QuoteStyle.None),"COUNTY STATE NAME"),
    9. #"Reordered Columns" = Table.ReorderColumns(#"Merged Columns",{"STATE", "COUNTY STATE NAME", "IDCENSUS", "CONUM", "CSA", "CBSA", "SCHLEV", "NCESID", "YRDATA", "V33", "TOTALREV", "TFEDREV", "C14", "C15", "C16", "C17", "C19", "B11", "C20", "C25", "C36", "B10", "B12", "B13", "TSTREV", "C01", "C04", "C05", "C06", "C07", "C08", "C09", "C10", "C11", "C12", "C13", "C24", "C35", "C38", "C39", "TLOCREV", "T02", "T06", "T09", "T15", "T40", "T99", "D11", "D23", "A07", "A08", "A09", "A11", "A13", "A15", "A20", "A40", "U11", "U22", "U30", "U50", "U97", "TOTALEXP", "TCURELSC", "TCURINST", "E13", "J13", "J12", "J14", "V91", "V92", "TCURSSVC", "E17", "E07", "E08", "E09", "V40", "V45", "V90", "V85", "J17", "J07", "J08", "J09", "J40", "J45", "J90", "J11", "J96", "TCUROTH", "E11", "V60", "V65", "J10", "J97", "NONELSEC", "V70", "V75", "V80", "J98", "TCAPOUT", "F12", "G15", "K09", "K10", "K11", "J99", "L12", "M12", "Q11", "I86", "Z32", "Z33", "V11", "V13", "V15", "V17", "V21", "V23", "V37", "V29", "Z34", "V10", "V12", "V14", "V16", "V18", "V22", "V24", "V38", "V30", "V32", "\_19H", "\_21F", "\_31F", "\_41F", "\_61V", "\_66V", "W01", "W31", "W61"})
    10. in
    11. #"Reordered Columns"

Kaggle\_School\_Data

1. Source: <https://www.kaggle.com/noriuk/us-education-datasets-unification-project?select=states_all_extended.csv>
2. Remove any samples that have null values for any feature.
   1. df = pd.read\_csv('states\_all.csv') : Load raw csv into panda frame.
   2. df= df.dropna() : Remove any rows with null values.