Project title

Report

Introduction and data

- Here are three potential research questions:

    1. Which three variables have the greatest impact on a college's ranking?

- These questions are important because students applying to college trust these rankings and weigh them into their college decisions. Due to large impact that a college has on a student's life, it is important to know where these ranking come from and what they actually measure.

- In general, we want to examine how different variables affect a school's ranking on the Niche College Ranking List. We plan to look at variables that are typically thought to influence school rank such as average SAT score and acceptance rate, but we also want to look at variables that aren't typically thought of such as geographic region or endowment size.

  We hypothesize that: ACT/SAT test score, acceptance rate, and median earnings 10 years after graduation will have the biggest impact on college rank.

- Here are the types of variables involved in our research questions:

    - Categorical (nominal):

        - geographic region
        - accreditor institution
        - control (public vs. private)
        - Carnegie classification

    - Numerical (continuous):

        - admission rate

        - Total share of enrollment by racial groups
        - average net price
        - average cost of attendance
        - average faculty salary
        - percentage of undergraduates who receive a Pell grant
        - completion rate
        - average age of entry
        - share of female students
        - share of married students
        - share of first generation students
        - median & mean family income
        - median & mean earning 10 years after entry
        - Value of school’s endowment
    - Numerical (discrete)
        - mean (if there is one) and median ACT/SAT scores for sub-tests and composite
        - Enrollment of undergraduate certificate/degree-seeking students

Data Tidying

```
Rows: 500 Columns: 2
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): college
dbl (1): rank

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.
```

```
Rows: 500
Columns: 2
$ college <chr> "Massachusetts Institute of Technology", "Stanford University"~
$ rank    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,~
```

```
Rows: 6681 Columns: 63
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (60): OPEID, OPEID6, INSTNM, CITY, STABBR, ZIP, ACCREDAGENCY, LATITUDE, ...
dbl  (3): UNITID, REGION, CONTROL

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.
```

```
Rows: 6,681
Columns: 63
$ UNITID            <dbl> 100654, 100663, 100690, 100706, 100724, 100751, 100760~
$ OPEID             <chr> "100200", "105200", "2503400", "105500", "100500", "10~
$ OPEID6            <chr> "1002", "1052", "25034", "1055", "1005", "1051", "1007~
$ INSTNM            <chr> "Alabama A & M University", "University of Alabama at ~
$ CITY              <chr> "Normal", "Birmingham", "Montgomery", "Huntsville", "M~
$ STABBR            <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", ~
$ ZIP               <chr> "35762", "35294-0110", "36117-3553", "35899", "36104-0~
$ ACCREDAGENCY      <chr> "Southern Association of Colleges and Schools Commissi~
$ LATITUDE          <chr> "34.783368", "33.505697", "32.362609", "34.724557", "3~
$ LONGITUDE         <chr> "-86.568502", "-86.799345", "-86.17401", "-86.640449",~
$ REGION            <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
$ CCBASIC           <chr> "18", "15", "20", "16", "19", "15", "2", "22", "18", "~
$ CONTROL           <dbl> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 3, 1, 1, 2, 1, 3, ~
$ ADM_RATE          <chr> "0.8965", "0.806", "NULL", "0.7711", "0.9888", "0.8039~
$ SATVR25           <chr> "430", "560", "NULL", "590", "438", "540", "NULL", "NU~
$ SATVR75           <chr> "520", "668", "NULL", "700", "531", "660", "NULL", "NU~
$ SATMT25           <chr> "410", "530", "NULL", "580", "406", "530", "NULL", "NU~
$ SATMT75           <chr> "500", "660", "NULL", "730", "518", "670", "NULL", "NU~
$ SATWR25           <chr> "370", "NULL", "NULL", "NULL", "NULL", "480", "NULL", ~
$ SATWR75           <chr> "457", "NULL", "NULL", "NULL", "NULL", "600", "NULL", ~
$ SATVRMID          <chr> "475", "614", "NULL", "645", "485", "600", "NULL", "NU~
$ SATMTMID          <chr> "455", "595", "NULL", "655", "462", "600", "NULL", "NU~
$ SATWRMID          <chr> "414", "NULL", "NULL", "NULL", "NULL", "540", "NULL", ~
$ ACTCM25           <chr> "15", "22", "NULL", "24", "14", "23", "NULL", "NULL", ~
$ ACTCM75           <chr> "20", "30", "NULL", "31", "20", "31", "NULL", "NULL", ~
$ ACTEN25           <chr> "14", "22", "NULL", "24", "14", "23", "NULL", "NULL", ~
$ ACTEN75           <chr> "20", "33", "NULL", "33", "20", "33", "NULL", "NULL", ~
$ ACTMT25           <chr> "15", "20", "NULL", "23", "14", "21", "NULL", "NULL", ~
$ ACTMT75           <chr> "18", "27", "NULL", "29", "20", "29", "NULL", "NULL", ~
$ ACTWR25           <chr> "NULL", "NULL", "NULL", "NULL", "NULL", "7", "NULL", "~
$ ACTWR75           <chr> "NULL", "NULL", "NULL", "NULL", "NULL", "8", "NULL", "~
$ ACTCMMID        <chr> "18", "26", "NULL", "28", "17", "27", "NULL", "NULL", ~
$ ACTENMID        <chr> "17", "28", "NULL", "29", "17", "28", "NULL", "NULL", ~
$ ACTMTMID        <chr> "17", "24", "NULL", "26", "17", "25", "NULL", "NULL", ~
$ ACTWRMID        <chr> "NULL", "NULL", "NULL", "NULL", "NULL", "8", "NULL", "~
$ SAT_AVG         <chr> "959", "1245", "NULL", "1300", "938", "1262", "NULL", ~
$ UGDS            <chr> "5090", "13549", "298", "7825", "3603", "30610", "994"~
$ UG              <chr> "NULL", "NULL", "NULL", "NULL", "NULL", "NULL", "NULL"~
$ UGDS_WHITE      <chr> "0.0159", "0.5496", "0.255", "0.7173", "0.0167", "0.76~
$ UGDS_BLACK      <chr> "0.9022", "0.2401", "0.6913", "0.0907", "0.9265", "0.1~
$ UGDS_HISP       <chr> "0.0116", "0.061", "0.0268", "0.0599", "0.013", "0.051~
$ UGDS_ASIAN      <chr> "0.0012", "0.0704", "0.0034", "0.0354", "0.0019", "0.0~
$ UGDS_AIAN       <chr> "0.0028", "0.0024", "0", "0.0083", "0.0017", "0.0033",~
$ UGDS_NHPI       <chr> "0.0008", "0.0004", "0", "0.001", "0.0017", "0.0008", ~
$ UGDS_2MOR       <chr> "0.0143", "0.0469", "0", "0.0431", "0.0119", "0.0359",~
$ UGDS_NRA        <chr> "0.0073", "0.0232", "0", "0.019", "0.0155", "0.0187", ~
$ UGDS_UNKN       <chr> "0.044", "0.0059", "0.0235", "0.0252", "0.0111", "0.00~
$ NPT4_PUB        <chr> "15529", "16530", "NULL", "17208", "19534", "20917", "~
$ NPT4_PRIV       <chr> "NULL", "NULL", "17618", "NULL", "NULL", "NULL", "NULL~
$ COSTT4_A        <chr> "23445", "25542", "20100", "24861", "21892", "30016", ~
$ COSTT4_P        <chr> "NULL", "NULL", "NULL", "NULL", "NULL", "NULL", "NULL"~
$ AVGFACSAL       <chr> "7599", "11380", "4545", "9697", "7194", "10349", "658~
$ PCTPELL         <chr> "0.7095", "0.3397", "0.7452", "0.2403", "0.7368", "0.1~
$ C150_4          <chr> "0.2866", "0.6117", "0.25", "0.5714", "0.3177", "0.721~
$ AGE_ENTRY       <chr> "20.28374137", "23.60797466", "33.6722973", "22.727919~
$ FEMALE          <chr> "0.564030132", "0.63909074", "0.648648649", "0.4763499~
$ MARRIED         <chr> "0.009102323", "0.105086641", "0.236486487", "0.100460~
$ FIRST_GEN       <chr> "0.365828092", "0.341223671", "0.5125", "0.310132159",~
$ FAMINC          <chr> "32362.82611", "51306.67431", "21079.47297", "61096.58~
$ MD_FAMINC       <chr> "23553", "34489", "15033.5", "44787", "22080.5", "6673~
$ MN_EARN_WNE_P10 <chr> "35500", "48400", "47600", "52000", "30600", "51600", ~
$ MD_EARN_WNE_P10 <chr> "36339", "46990", "37895", "54361", "32084", "52751", ~
$ ENDOWBEGIN      <chr> "NULL", "537349307", "174805", "77250279", "94536751",~
```

The Observations for `University of South Florida - Sarasota-Manatee` and `University of South Florida - St. Petersburg` have been dropped from the `colleges` data set due to their non-existence in the `us_dep_of_ed` data set. They were not present in the data set because these two universities were combined with the main University of South Florida campus. This is why the `colleges` data set only has 498 observations

Methodology

The methodology section should include visualizations and summary statistics relevant to your research question. You should also justify the choice of statistical method(s) used to answer your research question.

Prepare Data

1. Pick variables that make sense to focus on
2. Standardize all the continuous variables to z-scores, mean = 0, standard deviation = 1
3. Check correlation coefficients between all the the all the exponential variables. For the ones that have a $r > 0.8$, pick one to put in the model. We can create multiple models with different options.

Build model (Iterations)

1. Put all of our chosen variables into a model
2. Remove variables with the lowest slope coefficients until we get the best adjusted r-squared for the overall models.
3. Look at slope coefficients for variables in models for variables in the model. The one with the highest slope is the most significant.

Results

Showcase how you arrived at answers to your research question using the techniques we have learned in class (and beyond, if you're feeling adventurous).

Provide only the main results from your analysis. The goal is not to do an exhaustive data analysis (calculate every possible statistic and perform every possible procedure for all variables). Rather, you should demonstrate that you are proficient at asking meaningful questions and answering them using data, that you are skilled in interpreting and presenting results, and that you can accomplish these tasks using R. More is not better.