Week 12 - Final Data Project

STAT 420, Summer 2022, D. Unger

The goal of the project is to apply what you have learned in this course to analyze a data set of your choosing.

Timeline

1. A *roster* of your group members is due by **Monday, July 11, 11:59pm** (beginning of Week 9).

2. A *proposal* of your intended project is due by **Wednesday, July 20, 11:59pm** (middle of Week 10).

3. The *project report* is due by **Friday, August 5, 11:59pm** (end of Week 12). Late submissions will not be accepted.

4. The *peer evaluation* is due by **Friday, August 5, 11:59pm** (end of Week 12). Late submissions will not be accepted.

Project Guidelines

1. The final data project is a group project meant to be completed as a team. If you have not formed/joined a group by the roster deadline, you will be randomly assigned to one. There will be no single-member groups.

2. Select a dataset. This dataset might be relevant to research outside of this course, another field, or some other interest of yours. It will be helpful if you have a few questions in mind and that the data that you select would be relevant to those questions. You may not select a dataset that has been or is related to a dataset featured in class lessons or homework.

3. The dataset must be "substantially large", i.e., abiding by the following parameters: at least 2000 observations and roughly 10 variables. Of the variables at least one must be able to serve as a **numeric** response variable of interest. The other variables must be able to serve as explanatory

variables with at least 1 being categorical and at least 2 being continuous numeric.

4. You will use the data to construct a model for predicting values of a response variable using the predictor variables.

5. You will apply data analysis techniques which may include some of the following but are not limited to:

- Data cleaning as necessary to address observations with missing or extreme values.
- Multiple linear regression
- ANOVA
- Dummy variables
- Interaction
- Residual diagnostics
- Outlier diagnostics
- Transformations
- Polynomial regression
- Stepwise model selection
- Variable selection

6. Regarding final model selection

- There is not necessarily one, singular correct answer/model, but certainly some methods and models are more useful and would perform better than others depending on the data you choose.
- You do not necessarily have to use all predictors.
- You may use any methods we studied this semester to complete this task and provide evidence that your final choice of model is a good one.
- Some methods will be more useful than others for your data. Please only show tables/results/plots associated with leading you toward your final model. Don't include results that lead to a dead end, however you might choose to write a sentence or two to explain those other methods and why or how they failed.
- You may assume that the reader has knowledge of all STAT 420 concepts we've covered, but that they have no knowledge or background in the specific data problem you've chosen. You will still need to explain the rationale for your decision-making, and the report should be a standalone document.
- This is intentionally open-ended to see how you do without being given explicit steps, so have fun building it.

Potential Sources for Your Data

If you have any questions about whether your data meets the specifications of the project given above, don't hesitate to ask. If you plan to use data from another endeavor of yours, such as a research project, be sure to gain permission from the controlling authority first.

- Data.gov (http://www.data.gov/)

- Centers for Disease Control and Prevention (http://www.cdc.gov/datastatistics/)
- StatSci.org (http://www.statsci.org/datasets.html)
- NIST's Statistical Reference Datasets (StRD) (http://www.itl.nist.gov/div898/strd/)
- U.S. Census Bureau (http://www.census.gov/data.html)
- U.S. Bureau of Labor Statistics (http://www.bls.gov/data/)
- Kaggle (https://www.kaggle.com/)
- Inter-university Consortium for Political and Social Research (http://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp)
- Data Counts! (http://www.ssdan.net/datacounts)
- MathForum.org Data Library (http://mathforum.org/library/topics/data_sets/)
- StatLib Data Library (Data, Software and News from the Statistics Community) (http://lib.stat.cmu.edu/DASL/)
- Sports-Reference family of sites (http://www.sports-reference.com/)

If you find alternative sources for good datasets, please post about them in the forums!

Assignment Specifics

Roster Details

A *roster* of your group members is due by the above deadline.

- The Data Project Roster signup can be found at this link (https://forms.illinois.edu/sec/1598252800).
- You may only signup for group with students enrolled in the same section as you.
- For MCSDS students, groups will be comprised of 2-3 students.
- For Campus students, groups will be comprised of 3-4 students.
- Groups not meeting the minimum size may have more students added to their group.
- There will be no single-member groups.
- If you wish to be randomly assigned to a group with other students, you still need to fill out the form.
- There is a 5-point deduction from the project grade for late sign-ups.

Proposal Details

A *proposal* of your intended project is due by the above deadline.

A proposal of your intended project should include the following.

1. The names of the students who will be contributing to the group project.
2. A tentative title for the project.
3. Description of the data file (what they contain including number of variables and number of records). You do not necessarily have to list all the variables, but at least mention those of greatest importance.

4. Background information on the data sets, including specific citation of their source (so that I can also access it).
5. A brief statement of the business, science, research, or personal interest you have in the data set which you hope to explore. **6.** Evidence that the data can be loaded into `R`. Load the data, and print the first few values of the response variable as evidence.

Simply submit a single document outlining the proposal.

After review of the proposal, it will be evaluated in one of two ways.

- Approved (1) - Your group may proceed with your plans for the data and project.
- Pending (0) - I will provide suggestions, concerns, or needed information that must be addressed before the proposal will be approved.
  There is a 5-point deduction from the project grade for late proposal submissions.

Project Report Details

1. Access to the data. Even if the data is publically accessible and cited with a link in your project report, you must still provide the actual raw data file, uploaded with your project report. If you find that raw data files are too large to upload, contact me for an alternate option. It is your responsibility to have a way to provide the data to me well before the deadline and unsuccessful last minute attempts will result in deductions to the grade.

2. The .Rmd program file. This should include only that executable code pertinent to the summary report. Since you are uploading the data as well, the .Rmd file should be able to import it directly. That is, point-and-click or import wizard methods for accessing the data in `R` are not allowed.

3. A summary report in html format that includes the following.

It should begin with the title of the project. Do not include the names of the group members at the outset of the report.

The **introduction** section should relay what you are attempting to accomplish. It would include a statement of the business, science, research, or personal interest you have that leads to analyzing the data you've chosen. It should provide enough background to your work such that a reader would not need to load your data to understand your report. Like the simulation project, you can assume the reader is familiar with the course concepts, but not your data. Some things to consider:

- What is this data? Where did it come from? What are the variables? Why is it interesting to you?
- Why are you creating a model for this data? What is the goal of this model?

The **methods** section should contain the bulk of your "work." This section will contain the bulk of the `R` code that is used to generate the results. Your `R` code is not expected to be perfect idiomatic `R`, but it is expected to be understood by a reader without too much effort. Use RMarkdown and code comments to your advantage to explain your code if needed.

This section should contain any information about data preparation that is performed to the original data before modelling. Then you will apply methods seen in class, which may include some of the following but are not limited to:

- Multiple linear regression
- Dummy variables
- Interaction
- Residual diagnostics
- Outlier diagnostics
- Transformations
- Polynomial regression
- Model selection

*Your task is **not** to use as many methods as possible*. Your task is to use appropriate methods to find a good model that can correctly answer a question about the dataset, and then to communicate your result effectively. Some possible items to be discussed:

- Description of the original data file including description of all relevant variables.
- Description of additional data preparation that you performed.
- Description of the process you chose to follow.
- Narrative of your step-by-step decision making process throughout the analysis as you adjusted the model and attempted to validate model assumptions.

The **results** section should contain numerical or graphical summaries of your results. You should report a final model you have chosen. There is not necessarily one, singular correct model, but certainly some methods and models are better than others in certain situations. You may use any methods we studied this semester to complete this task, and provide evidence that your final choice of model is a good one. Some possible items to be discussed:

The **discussion** section should contain discussion of your results and should frame your results in the context of the data. How is your final model useful?

The **appendix** section should contain code and analysis that is used, but that would have otherwise cluttered the report or is not directly related to the choice of model. Do not simply dump code in here. Only utilize the appendix to supplement the primary focus of the report. The appendix should also conclude with the names of the group members.

Write in complete sentences and pay attention to grammar, spelling, readability and presentation. If you include a table or chart, make sure you say something about it. If you're not discussing a result, then it doesn't belong in your report.

Submit the following three items in a .zip file just as you do in homework assignments.

- your selected data,
- a .Rmd program file,
- and the project report (.html file)

Peer Evaluation Details

Individually, you will write a short review of each of your group members, including yourself. For each member, comment on:

- Which parts of the project were worked on by that member
- How well that member communicated with the team (Provide a score from 0 to 100 as well as written comments.)
- How well that member understood the course concepts (Provide a score from 0 to 100 as well as written comments.)
- Proportion of the project completed by that member (Provide a proportion from 0% to 100% as well as written comments.)
- Individually, you will submit a single file (.pdf preferred) that contains your review.

Your grade for this task will come from how well you evaluate your group members. It is more important that you honestly review your team than give each member good remarks.

Project Report Evaluation

The grading rubric for the final project is summarized by five criteria, each worth 20 points. Here is a breakdown of the points for the analysis report:

- Use of Statistical Methodology
  - Have you used the appropriate methods for your dataset? Have you applied them correctly?
- Interpretation of Results
  - Do you arrive at the correct statistical conclusions from the analyses you perform?
- Discussion of Results
  - Do you discuss your analysis results in the context of the data?
- Use of R Programming
  - Similar to HW expectation: Does your code perform the desired task? Is your code readable?
- Organization and Presentation
  - Similar to HW expectation: Is your report easy to read? Does it use RMarkdown well? Is it written in a manner such that a reader does not already need to be familiar with the data?

Maximum total points: 100

Frequently Asked Question

The popular question is, “How long should it be?” You’ll be creating an html file, so there’s no such thing as a page count, and I’m not concerned with that anyway. On one hand, you need to provide results and evidence to support your decisions, and you need to be thorough and diligent as you walk

through the steps of finding your best model. On the other hand, a well-crafted data analysis will utilize brevity and conciseness. If you have a point to make, get to it. If you find yourself writing things simply for the sake of padding the word-count or trying to impress with lots and lots of code, you're including the wrong things in the report.