STAT 506: Midterm Exam
Name:

Please turn in the exam to GitHub and include the R Markdown code, a Word or PDF file with output. You are welcome to turn in code and output for each question separately. Please verify that all of the code has compiled and the graphics look like you think they should on your files, you are welcome to upload image files directly if they look distorted in the Word or PDF file.

While the exam is open book, meaning you are free to use any resources from class, this is strictly an individual endeavor and **you should not discuss the problems with anyone outside the course instructor including class members.** You are welcome to use any resources, but please include references and/or acknowledgments of the sources you have used. **Note this may be different on the Comprehensive Exams.**

For full credit, include your code and graphics for each question and create neat output by using options like `kable()` for tables and writing results in line with R commands.

Much of this exam will focus on a dataset with housing prices in Washington (but D.C. this time).

```r
library(readr)
DC <- read_csv("https://raw.githubusercontent.com/STAT506/Data/master/DC.csv")
```

The dataset is a sample from the data available at https://www.kaggle.com/christophercorrea/dc-residential-properties.

The following variables are retained for use.

- BATHRM: Number of Full Bathrooms
- HF_BATHRM: Number of Half Bathrooms (no bathtub or shower)
- AC: Cooling
- BEDRM: Number of Bedrooms
- STORIES: Number of stories in primary dwelling
- PRICE: Price of most recent sale
- CNDTN: Condition
- LANDAREA: Land area of property in square feet
- FULLADDRESS: Full Street Address
- ASSESSMENT_NBHD: Neighborhood ID
- WARD: Ward (District is divided into eight wards, each with approximately 75,000 residents)
- QUADRANT: City quadrant (NE,SE,SW,NW)

1. (2-way ANOVA: 22 points)

a. (4 points)

Create a set of polished graphics that illustrate the difference in housing price as a function of Ward (see https://planning.dc.gov/whatsmyward for more details) and whether the unit has air conditioning.

b. (2 points)

Write out a 2-way ANOVA for a model that includes Ward and air conditioning with complete linear model notation.

c. (4 points)

Fit a 2-way ANOVA and consider whether the model should include an interaction term. Define each term in the model and provide evidence and size statements.

d. (2 points)

Describe and assess the assumptions for this model.

e. (4 points)

Create a plot(s) that contains the uncertainty for the mean price for a home with and without air conditioning in each ward.

f. (4 points)

Create a plot(s) that contains the uncertainty for the price for a home with and without air conditioning in each ward.

g. (2 points)

Summarize the results from your model without using statistical jargon.

2. (Hierarchical Model: 24 points)

a. (4 points)

Compare and contrast how hierarchical models and a 2-way ANOVA with interaction terms can model different relationships across groups.

b. (4 points)

Write out a hierarchical model using Ward as grouping variable and air conditioning (AC) and detail the assumptions for this model.

c. (4 points)

Fit a hierarchical model using Ward as a grouping variable and AC. Define each term in the model and provide evidence and effect statements.

d. (2 points)

Describe and assess the assumptions for this model.

e. (4 points)

Create a plot(s) that contains the uncertainty for the mean price for a home with and without air conditioning in each ward.

f. (4 points)

Create a plot(s) that contains the uncertainty for the price for a home with and without air conditioning in each ward.

g. (2 points)

Summarize your model without using statistical jargon.

3. (54 points)

For this question, write a complete report about predicting the cost of properties in Washington, D.C. When writing, consider writing a report for a real estate development company in Washington, D.C. You can assume they have a statistical background equivalent to be your classmates. Please turn in a separate PDF of your work.

As part of this analysis, create a test and training set to evaluate a hierarchical regression model (using neighborhood as a grouping variable) and other relevant variables in the model. Include all figures and tables in the document. Combine figures into multi-panel summaries if necessary.

The report should include the following five sections:

1. Introduction
2. Data
3. Statistical Procedures
4. Results
5. Conclusion

A rubric is provided below for grading purposes the total score will be scaled to account for 56 points toward the final exam. Unless otherwise specified, each item is worth 4 points.

1. **No Credit**: Criterion was not addressed or was written in a way that was not understandable.
2. **Beginning**: Ideas are not clear and supporting ideas are not presented.
3. **Developing**: Ideas are identified but not well supported and developed or are minimally supported and developed.
4. **Advanced**: Ideas are clearly identified and are adequately supported and developed.

Report generalities:

- Spelling, grammar, writing clarity (8 points)
- Paragraphs, section labels, Length, Double spaced
- Appendix with complete code
- Acknowledgements and Citations for papers/packages other resources
- Proper use of statistical methods (12 points)

Introduction:

- Report motivation
- Sample size(s)
- Data source and study design
- Research question

Data:

- Variables with units and descriptive statistics
- Data Visualization (8 points)

Statistical Procedures (Hierarchical Model):

- Model Description / Justification (8 points)
- Discuss/assess assumptions
- Residuals: visual exploration

Results:

- Concise evidence statement for the hierarchical model
- Summarize patterns/“size” for the hierarchical model
- Summarize the predictive results for the hierarchical model

Discussion:

- Discussion about the predictive and explanatory performance of the model
- Scope of Inference: how can the results be generalized?