Homework 2

Due Thursday, September 30 at 11:59pm CST on Moodle (https://moodle.macalester.edu/mod/assign/view.php?id=27980)

Deliverables: Please use this template (template_rmds/hw2.Rmd) to knit an HTML document. Convert this HTML document to a PDF by opening the HTML document in your web browser. *Print* the document (Ctrl/Cmd-P) and change the destination to "Save as PDF". Submit this one PDF to Moodle.

Alternatively, you may knit your Rmd directly to PDF if you have LaTeX installed.

Project Work

Instructions

Goal: Begin an analysis of your dataset to answer your **regression** research question.

Collaboration: Form a team (2-3 members) for the project and this part can be done as a team. Only one team member should submit a Project Work section. Make sure you include the full names of all of the members in your write up.

Data cleaning: If your dataset requires any cleaning (e.g., merging datasets, creation of new variables), first consult the R Resources page (r-resources.html) to see if your questions are answered there. If not, post on the #rcode-questions channel in our Slack workspace to ask for help. *Please ask for help early and regularly* to avoid stressful workloads.

Required Analyses

1. **Initial investigation: ignoring nonlinearity (for now)**
    a. Use ordinary least squares (OLS) by using the `lm` engine and LASSO (`glmnet` engine) to build a series of initial regression models for your quantitative outcome as a function of the predictors of interest. (As part of data cleaning, exclude any variables that you don't want to consider as predictors.)
        - You'll need two model specifications, `lm_spec` and `lm_lasso_spec` (you'll need to tune this one).
    b. For each set of variables, you'll need a `recipe` with the `formula` , `data` , and pre-processing steps
        - You may want to have steps in your recipe that remove variables with near zero variance (`step_nzv()`), remove variables that are highly correlated with other variables (`step_corr()`), normalize all quantitative predictors (`step_normalize(all_numeric_predictors())`) and add indicator variables for any categorical variables (`step_dummy(all_nominal_predictors())`).

  - These models should not include any transformations to deal with nonlinearity. You’ll explore this in the next investigation.

c. Estimate the test performance of the models using CV. Report and interpret (with units) the CV metric estimates along with a measure of uncertainty in the estimate (`std_error` is readily available when you used `collect_metrics(summarize=TRUE)`).
  - Compare estimated test performance across the models. Which models(s) might you prefer?

d. Use residual plots to evaluate whether some quantitative predictors might be better modeled with nonlinear relationships.

e. Which variables do you think are the most important predictors of your quantitative outcome? Justify your answer. Do the methods you’ve applied reach consensus on which variables are most important? What insights are expected? Surprising?
  - Note that if some (but not all) of the indicator terms for a categorical predictor are selected in the final models, the whole predictor should be treated as selected.

Your Work

a & b.

```
# library statements
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
library(broom)
library(ggplot2)
library(tidymodels)
```

```
## Registered S3 method overwritten by 'tune':
##   method                   from
##   required_pkgs.model_spec parsnip
```

```
## ── Attaching packages ────────────────────────────────────── tidymodels 0.1.3 ──
## ✓ dials        0.0.9      ✓ tibble       3.1.4
## ✓ infer        1.0.0      ✓ tidyr        1.1.3
## ✓ modeldata    0.1.1      ✓ tune         0.1.6
## ✓ parsnip      0.1.7      ✓ workflows    0.2.3
## ✓ purrr        0.3.4      ✓ workflowsets 0.1.0
## ✓ recipes      0.1.16     ✓ yardstick    0.0.8
## ✓ rsample      0.1.0
```

```
## ── Conflicts ──────────────────────────────────────── tidymodels_conflicts() ──
## x purrr::discard()  masks scales::discard()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## • Use tidymodels_prefer() to resolve common conflicts.
```

```
tidymodels_prefer()
# read in data
NBAstats <- read_csv("Seasons_Stats.csv")
```

```
## New names:
## * `` -> ...1
```

```
## Rows: 24691 Columns: 53
```

```
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (3): Player, Pos, Tm
## dbl (48): ...1, Year, Age, G, GS, MP, PER, TS%, 3PAr, FTr, ORB%, DRB%, TRB%,...
## lgl  (2): blanl, blank2
```

```
##
## 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.
```

```
# data cleaning
NBAstatsWith3 <- NBAstats %>%
    filter(Year >= 1980)

NBAstats3Min <- NBAstatsWith3 %>%
    filter(MP >= 1300)

NBAstats3MinTrade <- NBAstats3Min %>%
    group_by(Year, Player) %>%
    mutate(num_entry = n()) %>%
    ungroup() %>%
    filter(num_entry ==1 | (num_entry > 1 & Tm == 'TOT')) %>%
    select(-blanl) %>%
    select(-blank2) %>%
    select(-GS) %>%
    select(-OWS) %>%
    select(-DWS) %>%
    select(-WS)

NBAstats3MinTrade[is.na(NBAstats3MinTrade)] = 0
```

```
# creation of cv folds
set.seed(123)
NBAstats_cv10 <- vfold_cv(NBAstats3MinTrade, v = 10)
```

```
# model spec
lm_lasso_spec <-
  linear_reg() %>%
  set_args(mixture = 1, penalty = 0) %>% ## mixture = 1 indicates Lasso
  set_engine(engine = 'glmnet') %>% #note we are using a different engine
  set_mode('regression')
```

```
# recipes & workflows

NBAstats_rec <- recipe( `WS/48` ~ . , data = NBAstats3MinTrade) %>%
  update_role(Player, new_role = "Player") %>% # we don't want to use ID as predictor
    update_role(Tm, new_role = "Tm") %>%
    update_role(Pos, new_role = "Pos") %>%
    update_role(`...1`, new_role = "...1") %>%
    update_role(Year, new_role = "Year") %>%
  step_novel(all_nominal_predictors()) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_nzv(all_predictors()) %>%
    step_corr(all_predictors()) %>%
  step_normalize(all_numeric_predictors())  # important step for LASSO

lasso_wf_NBAstats <- workflow() %>%
  add_recipe(NBAstats_rec) %>%
  add_model(lm_lasso_spec)

lasso_fit_NBAstats <- lasso_wf_NBAstats %>%
  fit(data = NBAstats3MinTrade) # Fit to entire data set (for now)

tidy(lasso_fit_NBAstats) # penalty = 0; equivalent to lm
```

```
## Loading required package: Matrix
```

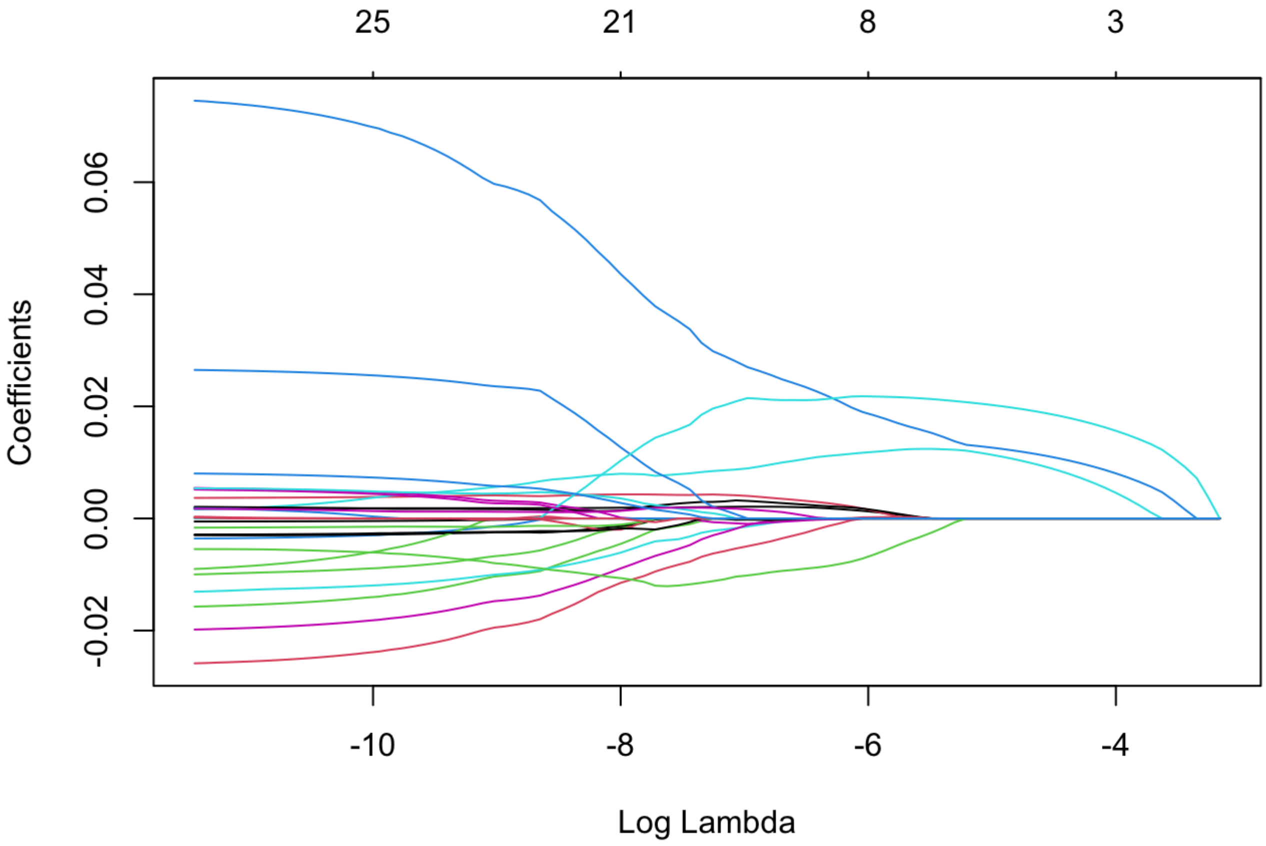
```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-2
```

```
## # A tibble: 29 × 3
##    term         estimate penalty
##    <chr>           <dbl>   <dbl>
##  1 (Intercept)  0.107          0
##  2 Age          0.00182        0
##  3 G            0.00367        0
##  4 MP          -0.00998        0
##  5 PER          0.0745         0
##  6 TS%          0.00160        0
##  7 3PAr         0.00544        0
##  8 FTr          0.00207        0
##  9 ORB%         0.000149       0
## 10 DRB%        -0.0157         0
## # … with 19 more rows
```

```
plot(lasso_fit_NBAstats %>% extract_fit_parsnip() %>% pluck('fit'), # way to get the ori
ginal glmnet output
     xvar = "lambda") # glmnet fits the model with a variety of lambda penalty values
```



```
# fit & tune models
lm_lasso_spec_tune <-
  linear_reg() %>%
  set_args(mixture = 1, penalty = tune()) %>% ## tune() indicates that we will try a variety of values
  set_engine(engine = 'glmnet') %>%
  set_mode('regression')

lasso_wf_NBAstats <- workflow() %>%
  add_recipe(NBAstats_rec) %>%
  add_model(lm_lasso_spec_tune)

penalty_grid <- grid_regular(
  penalty(range = c(-3, 1)), #log10 transformed
  levels = 30)

tune_output <- tune_grid( # new function for tuning parameters
  lasso_wf_NBAstats, # workflow
  resamples = NBAstats_cv10, # cv folds
  metrics = metric_set(rmse, mae),
  grid = penalty_grid # penalty grid defined above
)
```

c.

```
#  calculate/collect CV metrics
collect_metrics(tune_output)%>%
    select(penalty, rmse = mean, mae = mean)
```

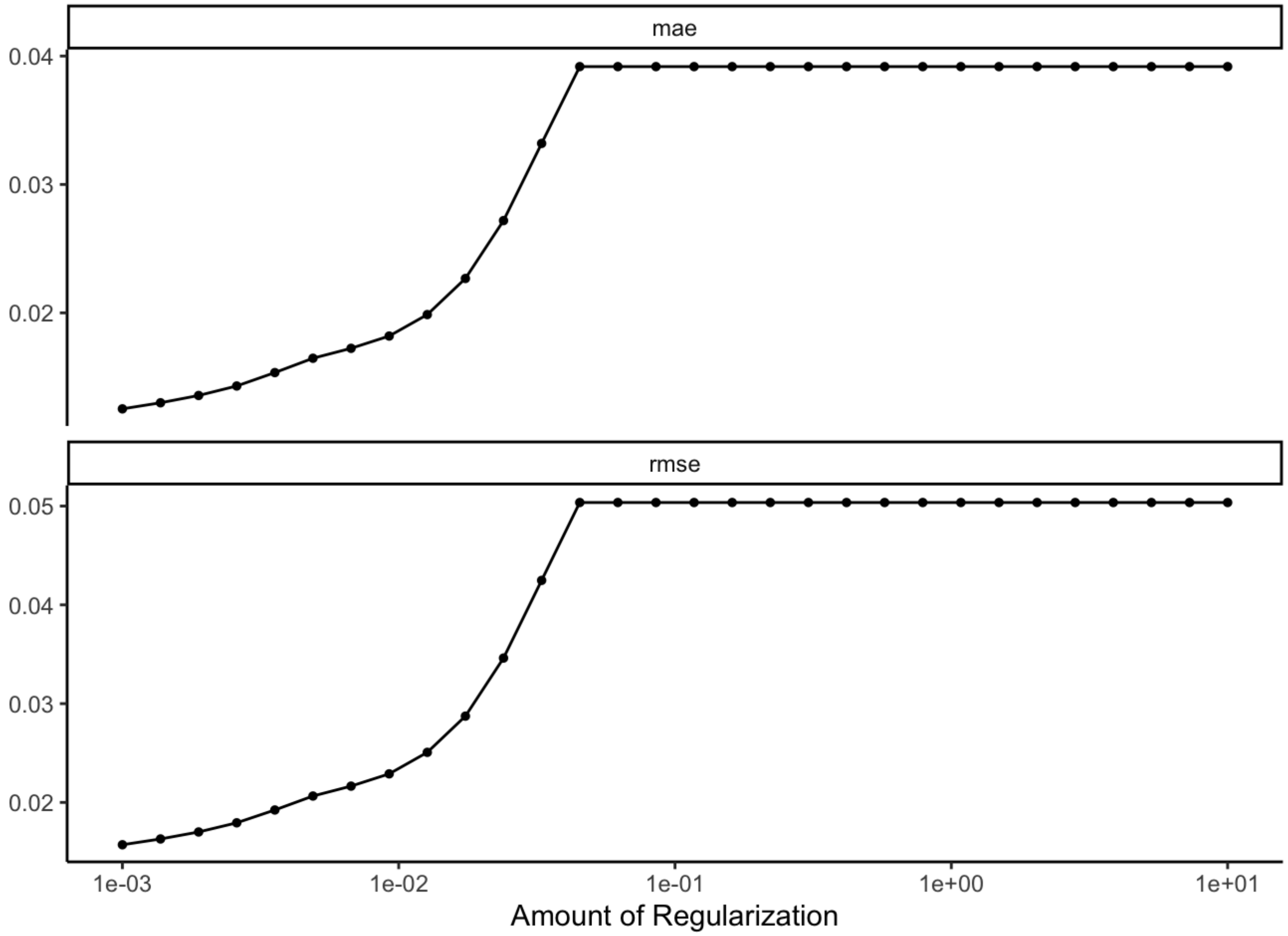
```
## # A tibble: 60 × 3
##    penalty   rmse    mae
##      <dbl>  <dbl>  <dbl>
##  1 0.001   0.0125 0.0125
##  2 0.001   0.0157 0.0157
##  3 0.00137 0.0130 0.0130
##  4 0.00137 0.0163 0.0163
##  5 0.00189 0.0136 0.0136
##  6 0.00189 0.0170 0.0170
##  7 0.00259 0.0143 0.0143
##  8 0.00259 0.0179 0.0179
##  9 0.00356 0.0153 0.0153
## 10 0.00356 0.0192 0.0192
## # … with 50 more rows
```

```
best_penalty <- select_best(tune_output, metric = 'rmse') # choose best penalty value

NBAstats_final_wk <- finalize_workflow(lasso_wf_NBAstats, best_penalty) # incorporates penalty value to workflow

NBAstats_final_fit <- fit(NBAstats_final_wk, data = NBAstats3MinTrade)

tidy(NBAstats_final_fit)
```

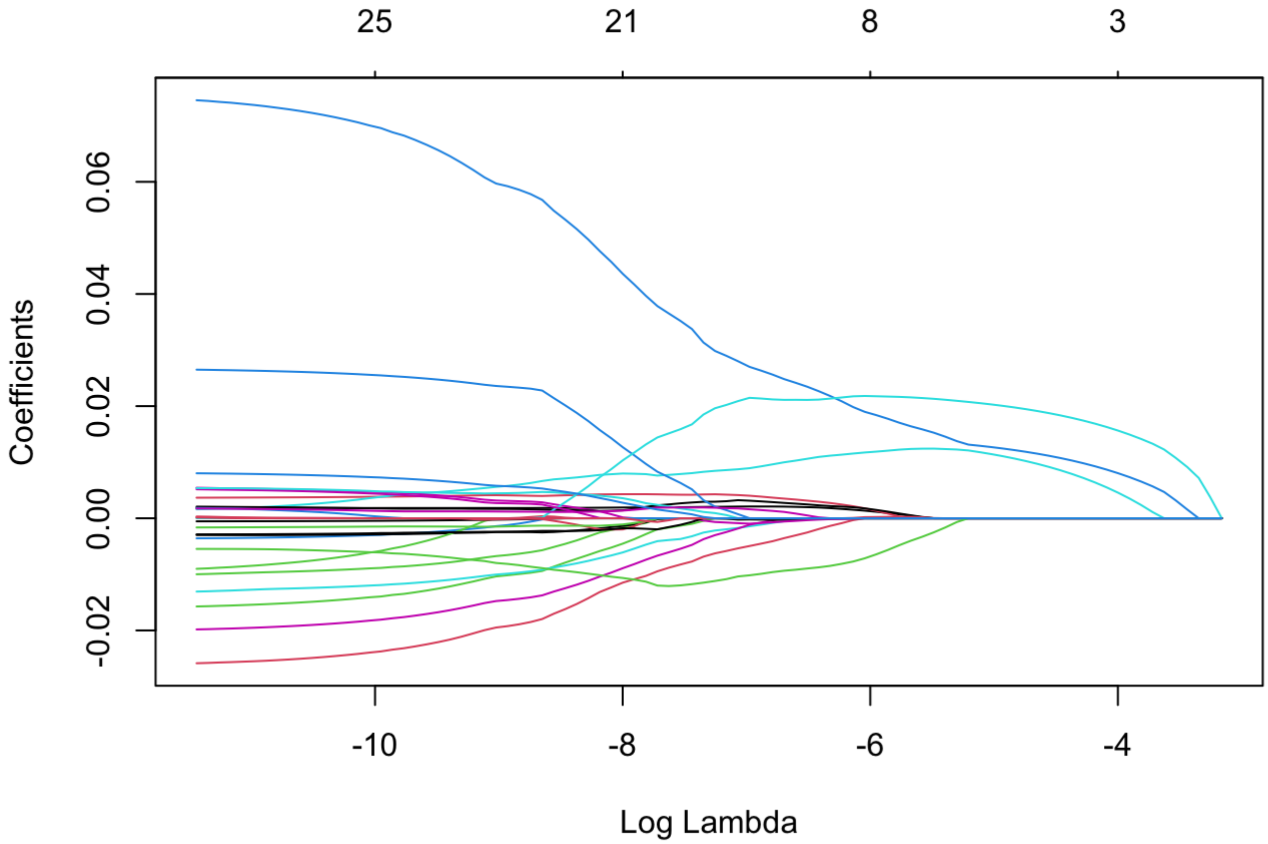
```
## # A tibble: 29 × 3
##    term           estimate penalty
##    <chr>             <dbl>   <dbl>
##  1 (Intercept)  0.107        0.001
##  2 Age          0.00210      0.001
##  3 G            0.00391      0.001
##  4 MP           0            0.001
##  5 PER          0.0265       0.001
##  6 TS%          0.00917      0.001
##  7 3PAr        -0.00000765   0.001
##  8 FTr          0.00305      0.001
##  9 ORB%         0            0.001
## 10 DRB%         0            0.001
## # … with 19 more rows
```

d.

```
# visual residuals
autoplot(tune_output) + theme_classic()
```



```
plot(NBAstats_final_fit %>% extract_fit_parsnip() %>% pluck('fit'), # way to get the original glmnet output
     xvar = "lambda") # glmnet fits the model with a variety of lambda penalty values
```



e.

2. **Summarize investigations**
   - Decide on an overall best model based on your investigations so far. To do this, make clear your analysis goals. Predictive accuracy? Interpretability? A combination of both?

Our best model to predict Win Share per 48 minutes would include the following predictors: Age, Games, Player Efficiency Rating, True Shooting Percentage, Three Point Attempt Rate, Free Throw Rate, Usage Percentage, Steal Percentage, Block Percentage, Turnover Percentage, Usage Percentage, Defensive Box Plus Minus, Box Plus Minus, Three Pointers, Free Throw Percentage, and Turnovers. When making this model we hoped to get a combination of both having a good predictive accuracy while also being able to be interpreted for future data, but we found ourselves having to sacrifice the suggested accuracy, as the residuals showed using a smaller penalty value and more predictors, to reduce the number of predictors and retain some interpretability.

3. **Societal impact**
   - Are there any harms that may come from your analyses and/or how the data were collected?
   - What cautions do you want to keep in mind when communicating your work?

There aren’t really any harms with our dataset as it is just about basketball and came from basketball reference, the officially used statkeepers of the NBA. The only cautions we should take is ensuring that it is knows that some stats are having both their version as a total and their versions as a percentage used, for example turnovers, while others may only have a percentage used as a predictor or a total like three pointers.

Portfolio Work

Length requirements: Detailed for each section below.

Organization: To help the instructor and preceptors grade, please organize your document with clear section headers and start new pages for each method. Thank you!

Deliverables: Continue writing your responses in the same Google Doc that you set up for Homework 1. Include that URL for the Google Doc in your submission.

Note: Some prompts below may seem very open-ended. This is intentional. Crafting good responses requires looking back through our material to organize the concepts in a coherent, thematic way, which is extremely useful for your learning.

Revisions:

- Make any revisions desired to previous concepts. **Important note:** When making revisions, please change from "editing" to "suggesting" so that we can easily see what you've added to the document since we gave feedback (we will "accept" the changes when we give feedback). If you don't do this, we won't know to reread that section and give new feedback.
- General guideance for past homeworks will be available on Moodle (under the Solutions section). Look at these to guide your revisions. You can always ask for guidance in office hours as well.

New concepts to address:

- **Subset selection:**
    - Algorithmic understanding: Look at Conceptual exercise 1, parts (a) and (b) in ISLR Section 6.8. **What are the aspects of the subset selection algorithm(s) that are essential to answering these questions, and why?** (Note: you'll have to try to answer the ISLR questions to respond to this prompt, but the focus of your writing should be on the question in bold here.)
    - Bias-variance tradeoff: What "tuning parameters" control the performance of this method? How do low/high values of the tuning parameters relate to bias and variance of the learned model? (3 sentences max.)
    - Parametric / nonparametric: Where (roughly) does this method fall on the parametric-nonparametric spectrum, and why? (3 sentences max.)
    - Scaling of variables: Does the scale on which variables are measured matter for the performance of this algorithm? Why or why not? If scale does matter, how should this be addressed when using this method? (3 sentences max.)
    - Computational time: What computational time considerations are relevant for this method (how long the algorithms take to run)?
    - Interpretation of output: What parts of the algorithm output have useful interpretations, and what are those interpretations? **Focus on output that allows us to measure variable importance. How do the algorithms/output allow us to learn about variable importance?**
- **LASSO:**

  - Algorithmic understanding: Come up with your own analogy for explaining how the penalized least squares criterion works.
  - Bias-variance tradeoff: What tuning parameters control the performance of this method? How do low/high values of the tuning parameters relate to bias and variance of the learned model? (3 sentences max.)
  - Parametric / nonparametric: Where (roughly) does this method fall on the parametric-nonparametric spectrum, and why? (3 sentences max.)
  - Scaling of variables: Does the scale on which variables are measured matter for the performance of this algorithm? Why or why not? If scale does matter, how should this be addressed when using this method? (3 sentences max.)
  - Computational time: What computational time considerations are relevant for this method (how long the algorithms take to run)?
  - Interpretation of output: What parts of the algorithm output have useful interpretations, and what are those interpretations? **Focus on output that allows us to measure variable importance. How do the algorithms/output allow us to learn about variable importance?**
- **KNN:**
  - Algorithmic understanding: Draw and annotate pictures that show how the KNN (K = 2) regression algorithm would work for a test case in a 2 quantitative predictor setting. Also explain how the curse of dimensionality affects KNN performance. (5 sentences max.)
  - Bias-variance tradeoff: What tuning parameters control the performance of this method? How do low/high values of the tuning parameters relate to bias and variance of the learned model? (3 sentences max.)
  - Parametric / nonparametric: Where (roughly) does this method fall on the parametric-nonparametric spectrum, and why? (3 sentences max.)
  - Scaling of variables: Does the scale on which variables are measured matter for the performance of this algorithm? Why or why not? If scale does matter, how should this be addressed when using this method? (3 sentences max.)
  - Computational time: The KNN algorithm is often called a “lazy” learner. Discuss how this relates to the model training process and the computations that must be performed when predicting on a new test case. (3 sentences max.)
  - Interpretation of output: The “lazy” learner feature of KNN in relation to model training affects the interpretability of output. How? (3 sentences max.)

Joe Margolis Portfolio

https://docs.google.com/document/d/12qm3A6qFtpUczJQmXUdNDhEnJfgyGIt_4j-niHFE348/edit?usp=sharing (https://docs.google.com/document/d/12qm3A6qFtpUczJQmXUdNDhEnJfgyGIt_4j-niHFE348/edit?usp=sharing)

Reflection

Ethics: Read the article Automated background checks are deciding who’s fit for a home (https://www.theverge.com/platform/amp/2019/2/1/18205174/automation-background-check-criminal-records-corelogic). Write a short (roughly 250 words), thoughtful response about the ideas that the article brings forth. What themes recur from last week’s article (on an old Amazon recruiting tool) or movie (Coded Bias)? What aspects are more particular to the context of equity in housing access?

```
This article shares one major underlying theme with the article I read last week on the
Amazon Recruiting tool, and it is a major issue with a lot of modern day uses of compute
r coding softwares. The issue is the widespread misconception that computers know all an
d that a calculator tool can tell more than a human, when in reality a computer only kno
ws as much as humans tell it to know, and even that becomes restricted by what coding la
nguages allow a human to tell it. This article highlights a coding software company call
ed CoreLogic which creates software that landlords use to vet potential tenants, particu
larly highlighting the software's ability to look up criminal history in applicants. The
way the article describes it, the software pretty simply just takes the name of the appl
icant and looks it up in a massive list of criminal and correctional facility databases,
and if there is a match the applicant fails the test and is likely rejected by the landl
ord.
This simple process of pass/fail for the applicants proves to be extremely harmful. One
of the ways, which was highlighted through the main storyline of the article through the
Arroyo family, is that the software only looks at past criminal history to try to predic
t the future, when there are many cases where the past is not telling of the future. In
this case for example, the applicant, who was docked on a crime lower than a misdemeano
r, another major flaw in the system, is physically unable to carry out a future crime. O
ther mentioned issues with the system included people with names matching criminals and
people with wiped records getting flagged. This is just more proof that in many ways com
puter software comes short on explaining small nuances in contextual environment and can
lead to great harm when relied on for major issues like housing.
```

Reflection: Write a short, thoughtful reflection about how things went this week. Feel free to use whichever prompts below resonate most with you, but don't feel limited to these prompts.

- How are class-related things going? Is there anything that you need from the instructor? What new strategies for watching videos, reading, reviewing, gaining insights from class work have you tried or would like to try?
- How is group work going? Did you try out any new collaboration strategies with your new group? How did they go?
- How is your work/life balance going? Did you try out any new activities or strategies for staying well? How did they go?

  This week has definitely been ramping up and things are starting to get very tough in all of my classes. I have begun to lock down a routine for getting all my classwork done, but with classwork ramping up the routines constantly need tweaking. Along with that I am starting to begin internship applications for this next summer which piles on more. The classwork itself is actually going well though, I am retaining the informationreally well I believe and getting good understanding of the material, but it certainly has been tough to feel like I am getting time to let everything sink in when I have to go right into other classwork after this course. I think a little bit of the stress might have come from not having a partner until this last Tuesday, so now that I can actually use the full two weeks to work on this homework things hopefully should feel a little easier.

Self-Assessment: Before turning in this assignment on Moodle, go to the individual rubric shared with you and complete the self-assessment for the general skills (top section). After "HW2:", assess yourself on each of the general skills. Do feel like you've grown in a particular area since HW1?

Assessing yourself is hard. We must practice this skill. These "grades" you give yourself are intended to have you stop and think about your learning as you grow and develop the general skills and deepen your understanding of the course topics. These grades do not map directly to a final grade.