**STAT 419 MIDTERM**

5 pts: 9 Qs (45 pts) 10 pts: 7 Qs (70 pts) 15 pts: 1 Q (15 pts) 20 pts: 1 Qs (20 pts) Cur Tot: 90

**BEFORE**

* Make sure HumanVerseWSU is working
* Remember to paste this information in your "console" before trying to Knit.  The database connection will not work without it...  \_SECRET\_database\_.txt

Sys.setenv(WSU\_SANDBOX\_HOST = "md5.mshaffer.com");

Sys.setenv(WSU\_SANDBOX\_DATABASE = "wsu\_sandbox\_db");

Sys.setenv(WSU\_SANDBOX\_USER = "wsu\_sandox\_user");

Sys.setenv(WSU\_SANDBOX\_PASSWD = "!WSUCougars");

**1. EDA Opinion (10 POINTS)**

* What is your opinion on this topic?
  + This is worth 10 points, a minimal answer should be at least 3 paragraphs.
  + Agreeing/Disagreeing with my opinion is not how you will be evaluated.
  + How well you express YOUR opinion is what is important.

**2. Basic Simulation (10 POINTS)**

* Code of Simulation
  + Pick a `set.seed` choice so the code is replicable. Verify that every time you run the commands, the data is not changing with the seed "you chose".
  + Use the functions `rnorm`, `runif` to simulate data.
  + Simulate `n=9999;` data for each.
  + Call `x.rnorm` the data for the first and `x.runif` the data for the second.
  + Plot a histogram `graphics::hist` and report the summary statistics ``base::summary` of each.
  + Then, plot them using `plot(x.rnorm, x.runif);`.
  + Finally, `plot(x.rnorm, sample(x.rnorm) );` and compare it to `plot(x.runif, sample(x.runif) );
* Describe `rnorm`, `runif
  + Describe what each function `rnorm` and `runif` does.
    - How are they similar?
    - How are they different?
  + What does the `sample` function do?
  + How was `plot(x.rnorm, x.runif);` different from `plot(x.rnorm, ( x.rnorm.sample = sample(x.rnorm) ) );` and `plot(x.runif, ( x.runif.sample = sample(x.runif) ) );`?
    - How would you describe the shape of each of these plots?

**3. “Easter-Egg" Simulation (5 POINTS)**

* There was an "Easter Egg" that related to setting the seed `set.seed` using `rbinom`.
* If you search in the BlackBoard discussion forum for `easter` you will see the discussion about August 25-27.
* In the "Easter Egg", the goal was to find a scenario using a specific `set.seed` that would simulate flipping a coin 100 times and getting one result (heads/tails) exactly 52 times.
* In this problem, the search criteria has changed. Simulate flipping a coin 1000 times and getting one result (heads/tails) exactly 555 times.
* You need to report 5 values for `set.seed` that achieves this objective. You can report more.
* You should explicitly have the code print `length(x)` where `x` is a vector of the values that meet the objective.

**Rolling the Dice**

* You have 3 dice.
* Each dice has the numbers `1:10` ... they are ten-sided die ("decader" die).
* Write the necessary `for-loops` to capture all possible outcomes of rolling the three dice at the same time.
* A dataframe `myrolls` should have three columns: `dice.1`, `dice.2`, `dice.3` plus a fourth column `roll.total` which is the sum `dice.1 + dice.2 + dice.3` of one iteration of the nested `for loop`.
* Report the dimensions `dim` of `myrolls`.
* Create a table `outcomes.table` that summarizes the counts of the `myrolls$roll.total`
* Transform the table to a dataframe `outcomes.df`. Name the columns: c("roll.total", "count");
* Report the sum of `outcome.df$count`
* Create a new column `outcomes.df$prob` (Probability) that determines the probability of that row given the total sum of the `count` column.
* Display the dataframe.

**Viewing a subset of data to answer a question**

* How many ways can I roll a 23 when I throw the dice at the same time?
* What is the probability that I roll a 23 on a single throw?

**4. Questions from the Dice simulation (10 POINTS)**

* Roll 23
  + How many ways can I roll a 23 when I throw the dice at the same time?
  + What is the probability that I roll a 23 on a single throw?
* Roll 12 or 22
  + What is the probability that I roll a 12 or a 22 on a single throw?
* Roll 26 or 29
  + What is the probability that I roll a 26 or a 29 on a single throw?
* Roll 3 once/twice in a row
  + What is the probability that I roll a 3 on a single throw?
  + What is the probability that I roll a 3 twice in a row?
    - First throw = 3 \*\*AND\*\* second throw = 3?
* Roll 12 or lower
  + What is the probability that I roll at most a 12 on a single throw?
  + That is, a 12 or lower ...

**INITIAL EXPLORATION OF REAL DATA**

* For each "search phrase", I go to Indeed.com and download the first page of results.
* From this first page, I grab the "total count"
* An example is shown in a screenshot, taken this week.
* <http://md5.mshaffer.com/WSU_STATS419/_images_/Big-data.png>

**Import "jobs" data**

* Run code to import the data `jobs`

**5. Histogram and Box Plot (5 POINTS)**

* Create a `hist` and `boxplot`
* Report `summary(jobs$job.count)`
* What does the histogram tell you about the data?
* What does the boxplot tell you about the data?
* What does `summary` tell you about the data?

**Histogram and Box Plot of SUBSET**

* What does the histogram tell you about the data?
* What does the boxplot tell you about the data?

**6. Trends in Relevant Subset (10 POINTS)**

* Run code chunk to start

**Initial Perspective**

* What is your initial "perspective" of this data, now that you see it?
* Why are the lines "parallel-ish"?
* What kind of a trend is that?
* Now comment on the line of data for "Data analysis".
  + How is it trending?
  + How does it compare to other Search-Query Words?
* What is your first perspective?

**Missing Data "Git" Week 40?**

* Is this data missing or did it just drop to zero that week?
* How does this relate to the other data (remember the idea of "continuity" in mathematics)?
* What should you do about it?

**Is "Microsoft Office" bigger than "C++"?**

* We are comparing two items, and these are generic ways of communicating such a comparison.
* In context of this data problem, a formalized form of the question would be something like:
  + "Utilizing job count for a given search query, determine if the query 'Microsoft Office' has a larger job count than the query 'C++'?"
    - This question will need to be formalized if we are trying to draw specific conclusions, but the initiation of analysis "which is bigger" allows us to understand what the data says or does not say, through exploration.
* To answer the question:
  + Mathematically, if two lines are parallel, and one is above the other, can we use \*\*distance\*\* to draw a conclusion?
  + Now, many times in statistics we deal with noise in the data, it is not "deterministic" but "stochastic" ... so we need to understand the variability.
  + Based on the data we see can we not use "parallel-line" logic to conclude that they are different? This is one dimension of EDA, use mathematics. ("mathematics")

**Tukey invented the boxplot as a nice EDA representation of the data.**

* What logical inference can we make about the distances between the boxplots and the fact that no data is overlapping.
* This is another dimension of EDA, use "distance" and the boxplot "IQR" to compare two elements.
* What conclusion would we make? ("boxplot")
* Can we conclude the data are different based on "mathematics" or "boxplot"?
* Would a formal "inferential statistical test" tell us something different than logical inference?
* How do you think "formal tests" were derived if not from "mathematics" and "boxplot" (EDA)?

**Is "Statistics" bigger than "Java"?**

* Use "mathematics" and "boxplot" and "ttest" to answer the question:
  + Is "Statistics" bigger than "Java"?

**What about "Data science" and "Big data"?**

* Use "mathematics" and "boxplot" and "ttest" to make a conclusion.
* Next, think carefully about the nature of the data. Is the "collection-approach" flawed to make a conclusion comparing job-counts of these specific keywords?
  + How likely is it that a single-job posting may have both keywords?
* This is an example where "data-integrity" knowledge would surpass the other three logical conclusions.
* This intuition requires an understanding of what mathematicians call "set theory". If I am doing an independent search on keywords, is it possible that one job would show up in multiple searches. That is, being counted twice or more.
* Intuition and logic would also allow us to conclude that our other comparisons are "very likely okay".
  + Why?
* What would be an improved approach to collecting the data that would allow me to more accurately compare these two keywords?
* This is representative of why exploratory data analysis is essential. It provides us insight into the domain and highlights the need for better data, if we can find it.

**COMPUTING DISTANCES**

* Notebook on collecting data from Wikipedia

**7. Data Provenance defined (10 POINTS)**

* Imagine you are preparing for a job interview.
* Write a 90-second blurb describing "what is data provenance" and "why it matters".
  + I would suggest the STAR(S) approach mentioned in one of the notebooks.
  + Reference the "Wikipedia" project as an example of how one can implement the features.
* Probably about 200-250 words (with the 90 second limit)

**Geospatial distances**

* Look at the 50 state capitals of America (USA)

**8. Distance from one input to multiple outputs (5 POINTS)**

* My Hometown "Columbia Falls, Montana" `cfalls`
  + Find all ZIP codes within 22 miles of Columbia Falls, MT `cfalls` (use lat/long provide from the Wikipedia lookup)...
  + build the bounding "box" and perform the post-hoc "radial distance" computations (as we did in the homework).
    - why is the box not a square, but a rectangle? ...
      * see `factor.lat` and `factor.long` in function `buildBoundingBoxFromRadiusAndGivenLatitudeLongitude`
    - critique the visualization ...
      * what do you like?
      * what would make it better?
* Your Hometown of something like it
  + Instead of `cfalls.info`, you do `hometown.info`
    - a location in the continental US of your choosing (not in Montana, Alaska, or Hawaii). [Graphing will not work for Alaska/Hawaii, Alaska has "boroughs" not counties.]
    - find the latitude/longitude of the location you have selected (how and where to look that up?)
    - Initially start with a radius of `13 miles`
    - When you run the code, note how many total "neighbors"; if it is less than 20; increase the "miles" so at least 20 results are returned.
    - In the end, you should select a location and radius that works for you. And its visualization also works.
    - Be certain to review and update the parameters before calling these functions.

**U.S. State Capitals (cities)**

* From Wikipedia, we grabbed one page that listed the 50 U.S. cities that are designated the capitals of each individual state in America (United States of America).

**Initial Plotting**

* Plot the data on `usmap` (ggplot2)
* Plot the data using maths `voronoi` (tripack)
* Plot the data on `map` (base)

**9. Comparing "Visualization Options" (5 POINTS)**

* Visually, which is the most appealing to you? Why?
* Functionally, which presents the data most effectively? Why?
* When we create visualizations, it is essential to portray the data accurately. For example, there are times when putting Alaska/Hawaii next to California might be appropriate, and other times it might not be.
* What is a one key factor that would determine this appropriateness?

**10. Building the distance matrix (5 POINTS)**

* Let's start with geo-spatial distances. I will do `distMeeus` and you will do `distHaversine`
  + Examining the data above, is there much difference between these two calculations?
  + What is the conceptual difference between these two calculations?
    - Try `?distMeeus` or `?distHaversine`
  + Are the results similar?
  + Is there a more accurate distance algorithm for geo-spatial calculations? If so, what is it?
* Let's now do `manhattan` and `euclidean` which are more common in the "statistical clustering" domain. I will do `euclidean`.
  + What is the difference between these two calculations?
  + Are the results similar?

**HIERARCHICAL clustering as a function of distance**

* Analogy of Family
* Clustering U.S. capital cities based on latitude, longitude
* Understanding the `cutree`

**11. Review one clustering tree (dendrogram) (10 POINTS)**

* Choose either `hclust.ward2.dist.euclidean` or `hclust.complete.dist.euclidean`
  + and review how the U.S. state capitals are clustered. [I commented out one form, so you will have to re-run if you want to select that one.]
* Comment on the "face validity" of this approach based on your understanding about how the U.S. regions are defined?
* Are the North/South Dakotas together?
* What about the North/South Carolinas?
* What about the Pacific Northwest?
* While living in Kentucky, some people called the area "Kentuckiana" meaning Kentucky/Indiana.
* Does that show up?
* Also note anything that seems peculiar.

**GENERIC clustering**

* Arbitrary Aggregation

**12. Movie Aggregation [Arbitrary] for Will and Denzel (5 POINTS)**

* # (1) populate cluster.arbitrary
* # (2) summarize how many movies live in each (table count)

**13. Movie Aggregation [Decile] for Will and Denzel (5 POINTS)**

* # use # stats::quantile(x, prob=seq(0.1,0.9,by=0.1), type=1 );
* # (1) how many NA's are there ... keep them NA's
* # (2) for the rest of the data, break it up into deciles
* # (3) $cluster.deciles for a given movie should be NA, 1, 2, 3, ... 10
* # (4) summarize how many movies live in each (table count)

**2.9 CENTROID clustering (k-means) as a function of distance**

* Introduction
* My recommendations
* WIKIPEDIA CLIMATE DATA
* Basic Background Research
* One Graph

**14. One Research Graph (5 POINTS)**

* What do you like about this graphic?
* What do you dislike?
* Is it aesthetically pleasing?
* Is it functional?
* We can achieve a side-by-side comparison using the function described below. The first city will be graphed on the left, the second city on the right.
  + What do you like about this graphic?
  + What do you dislike?
  + Are the y-axis the same scale? Are the visible gridlines for each the same?
  + What is the difference between rain and snow on the graphic? Was that a good approach? How would you have done it?

**15. One Publication Graph (5 POINTS)**

* Once you have a template built, it is rather easy to modify it. Here I changed the background map, and all of the data/features stay the same:
* <http://md5.mshaffer.com/WSU_STATS419/_EXAMPLES_/fiddle_usmap/world.html>

**Which Features to Include in the Analysis**

* If we want to cluster cities, which decisions seem best?
* Why?
* As you can see from the code below, you just comment out two options, and can quickly rerun the analysis.

**WHICH MONTHS & WHICH COLUMNS**

* Run code chunk

**To scale or not to scale, that is the question**

* So let's do some analysis with all of the data available to us. Most of the data is in Temperature, with ranges from -42 degrees Fahrenheit (Helena, Montana) to 122 (Phoenix, Arizona).
* The precipitation data (rain and snow) is measured in inches. So should we scale the data. The answer in PCA and orthogonal projections is absolutely
  + YES, but for `hclust` and `kmeans` is that always the case?
* You can make a choice below, and observe how it influences your answers.

**WHICH X**

* Run code chunk

**Perform `k-means` on All Climate Features**

* With the selected features let's perform k-means.
* Let's select k=12.
* I will select all of the features, but you can change that if you wish.

**Descriptive of Sample**

* Above, you are just analyzing the general shapes. Which ones are "fuller" circles? Why?
* Which ones are not very "full circles"?
* Why?

**Computation of Clusters/Centroids**

* Run code chunk

**Cluster Membership and Centroid Attributes**

* Run code chunk

**16. Summarize Findings (10 POINTS)**

* Identify which states share a common cluster.
* For a given cluster, what are its primary characteristics
* Summarize your k-means findings for 12 clusters.

**17. 2.10 Correlation (15 POINTS)**

* Describe the correlation of July in "Record high F (C)" to the other numeric factors printed above.
* intuitively, which months do you think correlate most with latitude for this data?
* which correlate the least?
* is the correlation always the same sign (positive/negative), or does it change?
  + You can use the dataframe output to do this analysis, or create your own subset
* Describe the correlation of January in "Record low F (C)" to the other numeric factors printed above.

**"So What" is DATA ANALYSIS?**

* Statistics
  + There are 5 elements mentioned: collection, organization, analysis, interpretation, and presentation of "data".
    - Are those equally weighted?
    - That is, should we devote 20% of our time to each of those?
  + Now, consider the "analysis" stage.
  + I have suggested there are two camps: exploratory and confirmatory data analysis.
    - Are those equally weighted?
  + That is, should we devote 50% of our time to each of those?
  + Now, in an "equally-likely" scenario, we would have.
* Data Analytics
* Importance of 'Data'
* Apprenticeship as Learning a Trade
* Tools of the Trade
  + These are not the tools of the trade, but hopefully, they introduce you to key tools of the trade. What exactly are tools of the trade?
  + You will have an opportunity to write a response below.
* Dimensional Reduction, an Axiomatic View
* Skills of the Trade
  + Can you acquire an appreciation for "data intimacy"?
  + Can you track and document how data is curated?
  + Can you track and document the analyses you perform? Can you recreate them? Do you have basic version-control protocols in place?
  + Can you view data from multiple perspectives and synthesize those perspectives to identify the central them of the data? Can you be objective? Can you try and identify objective metrics to enlighten your understanding about the essence of data?
  + Can you experiment with different visualizations in search of an optimal "one graphic" result? Do you have practice using various visualization tools? Can you comprehend which visualization tool is appropriate for messaging (communicating results) to a particular audience?
  + Can you communicate and defend your findings to a particular audience? Are your communications professional? Is the final work product both simple and comprehensive: simple in its summary findings and comprehensive in its ability to be replicated and audited as necessary.

**18. YOUR OPINION OF DATA ANALYTICS (20 POINTS)**

* (1) what proportion of "statistics" should be divided among: collection, organization, analysis, interpretation, and presentation of "data" ... providing a `barplot` of your opinion within your response would seem appropriate
* (2) what tools of the trade should you be acquiring from the core courses? how are you doing in that acquisition process (e.g., tool X is ... and right now I feel like my understanding/proficiency of tool X ... ) ...
* (3) utilize the provided `plot` script to place the coure-course categories on the proposed x-y graph related to analytics practice (Applied vs Theoretical) and care of data integrity (Great Care vs Little Care) ... Also place your personal assessment on the plot script provided
* (4) evaluate your skill-level on the six "skills of the trade": Emerging (Nascent), Developing (Adolescent), Mastering (Mature). explain your evaluation and include other important skills you believe are relevant that are not included
* (5) Any other comments you would like to share.
* Run code chunk
  + ########################### basic plot setup #####
  + ########################### you can add elements here #####
  + ############# TODO ###### ... maybe change color for each data point