STAT 530 – Midterm Exam – Fall 2022

Note: For this midterm exam, **you are not allowed to receive help from *anyone except me* on the exams.** For example, you may not talk to other students about the exam problems, and you may not look at other students' exams, code, graphs, etc. Violations of this policy may result in a 0 on the exam, an F for the course, and/or punishment by the USC Office of Academic Integrity. Work on these completely by yourself, just as you would on an in-class exam!

IMPORTANT: The following problems involve real data sets! The answers you get about the data may not be "perfect" or idealized. Your reports should address the questions of the clients in the best way you can, but it is NOT expected that every question the client wants addressed will have a perfectly clear answer. Your job is to explain and show what you have learned from the data, in no more than three pages (per problem) of clear writing (plus any relevant graphs and plots that you want to include at the end).

Furthermore, it is expected that each statistician (student) will make somewhat different conclusions or illuminate somewhat different aspects of the data set. For many of these questions, there is not only one right answer!

By the way, although the clients' questions are numbered, your report ideally will NOT be in the form of a numbered set of answers. It's better if your report addresses the questions in a more unified and flowing way, and you can arrange the flow of the report however you feel is appropriate. Make your report reflect good statistical work, but just as importantly, do your best to make it interesting to the reader! If the report is so dry and boring that the client's eyes glaze over when he or she tries to read it, then you haven't been a successful statistical communicator.

While you are working on the exam: You can ask me questions, but I may or may not give an answer. If it is a clarification question, I may answer that, but if it is about which approach to take on a problem or how to do a data analysis, I will likely not answer that, since this is a test and I want you to know that (or to look back through your notes and course resources for insight). If you have a question about R code that is not working: If it's basically correct except for a tiny thing causing an error, I may be able to point you in the right direction. But if it is something substantially wrong, I will just advise you to rethink your approach and look back at class examples.

Grading Scale:

Each problem will be worth 30 points, for a total of 60 points. For each problem, your report will be graded based on Writing, Analysis, and Context. For example:

Writing (out of 10 points): How organized, clearly written, comprehensible, and grammatically correct is the report? Would the client reading this report be confident that it was written by an educated, well-trained statistical scientist?

Analysis (out of 10 points): Were the graphs and data analyses appropriate for the problem? Were the analyses carried out correctly? Were your statistical conclusions about the data set sensible and clearly justified by numerical or graphical evidence?

Context (out of 10 points): Were the questions answered in terms of the variables of the data set? Although you are not an expert in the field as your client is, have you attempted to frame your conclusions and interpretations in a subject-matter context rather than treating the data as simply a meaningless set of numbers?

1. You are a young data scientist with an unlikely new client: a veteran cattle rancher who has just sold 76 of his bulls. This rancher (whose name is Jones but whom everyone calls Ol' Tex) has always run his business by hunches and rancher's intuition, but he's heard some of these young whippersnappers straight out of business school talking about "analytics" and "data-driven decision-making". Ol' Tex ain't so sure about these newfangled number-crunchers with their fancy computers ("I've been with the professors," says he, "and I'm not sure they like my looks"). But he does have some data on his bulls, and he's willing to give you a try. Here are the variables that he has observed or measured on the bulls:

```
Breed (1 = Angus, 5 = Hereford, 8 = Simental) --- this is purely categorical
and should NOT be treated as numeric
SalePr = selling price ($)
YrHgt = yearling height (when the bull was 1 year old) at shoulder (in)
FtFrBody = fat free body wt (lbs)
PrctFFB = percent fat-free body weight (%)
Frame = size scale (1 = small to 8 = large)
BkFat = back fat (in)
SaleHt = sale height at shoulder (in)
SaleWt = sale weight (lbs)
```

You walk into his room with your pencil in your hand, ready to do some calculations, and find him staring, with furrowed brow, at the rows and columns of measurements in front of him: "Something is happening here, and I don't know what it is," mutters Mr. Jones. It's your job as his data scientist to tell him what's happening with his data. Here are the specific questions he would like answered:

1. Are there particular bull(s) that are highly unusual in terms of the measured characteristics? If so, identify them.
2. Are there notable associations/relationships between some of the variables? (if so, describe them)
3. Is there a way to graphically represent the raw data for the 76 bulls (or any subset of them that is of interest to you) and draw conclusions about the data set from such a graph?
4. Are there similarities or differences among the breeds that can be seen and displayed by looking at the data?
5. Can we find a few indices that describe the variation in the data set using a lesser dimension than the original set of variables? If so, what are those indices? Is there a convenient interpretation of any of the indices?
6. Can we graphically display the data in a low number of dimensions using such indices? What conclusions about the bulls (individual bulls or groups of bulls) can you draw from such a graph?
7. What are any other potentially interesting aspects of the data set that may be gleaned for these data? Is there any useful information, e.g., for Tex’s future bull sales, could be gleaned from this data set?

When you write your report, make sure it's in understandable, clear English: Ol' Tex doesn't go for that overly formal academic-ese. On the other hand, he's no dummy and in his younger days it was well-known that he was very well-read --- he's been through all of F. Scott Fitzgerald's books --- so you don't have to dumb it down too much. But ultimately, what Ol' Tex wants is practical answers: As he explains to you, "We have a saying in the cow business: Give me some milk or else go home." Hearing your consultant's fee, he grumbles about statisticians, "Anyway they already expect you to all give a check..." Make sure Ol' Tex considers this money well spent by providing him a report that will hand him a ticket to success in his business.

You will type a report of NO MORE THAN 3 pages detailing your analysis of the data and your conclusions. Keep in mind that the report should be written for two audiences: Tex himself, who knows about bulls but is not an expert in statistics; and your supervisor, who will be judging you and deciding on your possible promotion based on the statistical competence of the report. Your report should be understandable and meaningful to both audiences.

You may include graphs that illustrate and/or support your findings. If so, put the graphs at the end of the report, and you may refer to them in the body of the report. (The graphs do not have to count as part of the 3-page length.) Do NOT include computer code within the main body of your report. You may include such code in an appendix if you wish (but it is unlikely that it will be read ... Ol' Tex ain't got no time for that).

The data for this problem are given at the link "Bulls data" on the course web page. Below that on the course web page is a link to some R code that may be helpful in reading in and managing the data.

2. You are working as a consulting statistician for an NBA basketball team that is preparing for roster planning and contract negotiations with several of its players. Data have been gathered on 176 players. The variables measured include several performance attributes of the players (data are from the 2022 season). The 15 variables measured on each player are: Age (player's age in years), GS (number of games the player started for the season), FGPct (percentage of field goals attempted that the player successfully made – this includes close shots and "long-distance" shots), X3PPct, (percentage of 3-point field goals attempted that the player successfully made – these are only "long-distance" shots), X3PPct, (percentage of 2-point field goals attempted that the player successfully made – these are only "closer-distance" shots), FTPct (percentage of free throws attempted that the player successfully made – these are unguarded medium-distance shots following an opponent's foul), MinGame (average number of minutes the player played per game), ORBMin, DRBMin (number of offensive rebounds per minute and number of defensive rebounds per minute; typically tall players have more chance to get rebounds, especially defensive rebounds, since tall players tend to stand closer to the basket on defense), ASTMin (number of assists per minute; typically quick ball-handlers have more chance to get assists), STLMin (number of steals per minute; typically quick players have more chance to get steals), BLKMin (number of blocked shots per minute; typically tall players have more chance to get blocks), TOVMin (number of turnovers per minute; typically players who often handle the ball have more chance to get turnovers), PFMin (number of personal fouls per minute), PTSMin (number of points scored per minute).

The questions that the team would like answered include:

1. Are there individual players who are highly unusual (in any way) based on the measured performance-related variables (those variables other than age, starts, and minutes per game)? If so, identify their names.
2. Are there notable associations/relationships between some of the variables? (if so, describe them)
3. Is there a way to graphically represent the raw data for the 176 players (or more realistically, for some small subset of the players that is interesting to you) and draw conclusions about the data set from such a graph?
4. Are there a small number of underlying characteristics of players that the observed variables might be connected to? If so, determine how many latent characteristics there seem to be in this set of variables. Also, try to interpret them the best you can, with the aid of statistical techniques.
5. Can we graphically display the data in a low number of dimensions using such latent traits? What conclusions about the players (individual players or groups of players) can you draw from such graph(s)?
6. What are any other potentially interesting aspects of the data set?

You will type a report of NO MORE than 3 pages detailing your analysis of the data and your conclusions. Keep in mind that the report should be written for two audiences: the basketball team's client, who has a sense for numbers but is not an expert in statistics; and your own supervisor at the statistical consulting company, who will be judging you and deciding on your possible promotion based on the statistical competency of the report. Your report should be understandable and meaningful to both audiences.

You may include graphs that illustrate and/or support your findings. If so, put the graphs at the end of the report, and you may refer to them in the body of the report. (The graphs do not have to count as part of the 3-page length.) Do NOT include computer code within the main body of your report. This will be incomprehensible to the client and would only annoy her. You may include such code in an appendix if you wish.

The data for this problem are given at the link “NBA Players Data 2022” on the course web page. Below that on the course web page is a link to some R code that may be helpful in reading in and managing the data.

Note that the R code creates some other variables like the players’ names; shortened versions of the players’ names; the players’ positions; and the players’ teams. You may or may not find these useful; you can use them or not, as you wish. Note that when a player’s team is “TOT” this means the player was on more than one team during the 2022 season and his statistics are his total statistics across the whole season. (For the players’ positions: PG=point guard, SG=shooting guard, SF=small forward, PF=power forward, C=center. Typically, the guards are the shorter players and the centers the tallest players, with forwards being medium height.) Again, this is not necessarily relevant, but it is there if you care to look at it.