Introduction to Data Science – DS UA 112
Capstone project

The purpose of this capstone project is to tie everything we learned in this class together. This might be challenging in the short term, but is consistently rated as being extremely valuable in the long run. The cover story is that you are working as a Data Scientist for an auction house that works with a major art gallery. You want to better understand this space with the help of Data Science. Historically, this domain was dominated by art historians, but is increasingly informed by data. This is where you – a budding Data Scientist – come in. Can you provide the value that justifies your rather high salary?

Mission command preamble: As in general, we won't tell you *how* to do something. That is up to you and your creative problem solving skills. However, we will pose the questions that you should answer by interrogating the data. Importantly, we do expect you to do this work yourself, so it reflects your intellectual contribution – not that of third parties. By doing this assignment, you certify that it indeed reflects your individual intellectual work.

Format: The project consist of your answers to 10 (equally-weighed, grade-wise) questions. Each answer *must* include some **text** (describing both what you *did* and what you *found,* i.e. the answer to the question), a **figure** that illustrates the findings and some **numbers** (e.g. test statistics, confidence intervals, p-values or the like). Please save it as a pdf document. This document should be 4-6 pages long (arbitrary font size and margins). About half a page per question is reasonable. In addition, open your document with a brief statement as to how you handled preprocessing (e.g. dimension reduction, data cleaning and data transformations), as this will apply to all answers. Include your name.

Academic integrity: You are expected to do this project by yourself, individually, so that we are able to determine a grade for you personally. There are enough degrees of freedom (e.g. how to clean the data, what variables to compare, aesthetic choices in the figures, etc.) that no two reports will be alike. We'll be on the lookout for suspicious similarities, so please refrain from collaborating.
To prevent cheating (please don't do this – it is easily detected), it is very important that you – at the beginning of the code file – seed the random number generator with your N-number. That way, the correct answers will be keyed to your own solution (as this matters, e.g. for the specific train/test split or bootstrapping). As N-numbers are unique, this will also protect your work from plagiarism.
Failure to seed the RNG in this way will also result in the loss of grade points.

Deliverables: Upload two files to the Brightspace portal by the due date in the sittyba:
*A pdf (the "project report") that contains your answers to the 10 questions, as well as an introductory paragraph about preprocessing.
*A .py file with the code that performed the data analysis and created the figures.

We do wish you all the best in executing on these instructions. We aimed at an optimal balance between specificity and implementation leeway, while still allowing us to grade the projects in a consistent and fair manner.
Everything should be doable from what was covered in this course.
If you take this project seriously and do a quality job, you can easily use it as an item in your DS portfolio. Former students told us that they secured internships and even jobs by well executed capstone projects that impressed recruiters and interviewers.

Description of dataset: This dataset consists of data from 300 users and 91 art pieces.

The art pieces are described in the file "theArt.csv". Here is how this file is structured:
1st row: Headers (e.g. number, title, etc.)
Rows 2-92: Information about the 91 individual art piece
Column 1: "Number" (the ID number of the art piece)
Column 2: Artist
Column 3: Title
Column 4: Artistic style
Column 5: Year of completion
Column 6: Type code – 1 = classical art, 2 = modern art, 3 = non-human art
Column 7: computer or animal code – 0 = human, 1 = computer generated art, 2 = animal art
Column 8: Intentionally created? 0 = no, 1 = yes
You can also take a look at the actual art by looking at the files in the "artPieces" folder on Brightspace.

The user data is contained in the file "theData.csv". Here is how this file is structured:
Rows 1-300: Responses from individual users
Columns 1-91: Preference ratings (liking) of the 91 art pieces. The column number in this file corresponds to the number of the art piece in column 1 of "theArt.csv" file described above. For instance, ratings of art piece 27 ("the woman at the window") is in column 27. Numbers represent preference ratings from 1 ("hate it") to 7 ("love it").
Columns 92-182: "Energy" ratings of the same 91 art pieces (in the same order as the preference ratings above). Numbers represent ratings from 1 ("it calms me down a lot") to 7 ("it agitates me a lot").
Columns 183-194: "Dark" personality traits. Numbers represent how much a user agrees with a statement, from 1 (strongly disagree) to 5 (strongly agree). Here are the 12 statements, in column order:

1	I tend to manipulate others to get my way
2	I have used deceit or lied to get my way
3	I have used flattery to get my way
4	I tend to exploit others towards my own end
5	I tend to lack remorse
6	I tend to be unconcerned with the morality of my actions
7	I can be callous or insensitive
8	I tend to be cynical
9	I tend to want others to admire me
10	I tend to want others to pay attention to me
11	I tend to seek prestige and status
12	I tend to expect favors from others

Columns 195-205: Action preferences. Numbers represent how much a user agrees with a statement, from 1 (strongly disagree) to 5 (strongly agree). Here are the 11 actions, in column order:

1 I like to play board games
2 I like to play role playing (e.g. D&D) games
3 I like to play video games
4 I like to do yoga
5 I like to meditate
6 I like to take walks in the forest
7 I like to take walks on the beach
8 I like to hike
9 I like to ski
10 I like to do paintball
11 I like amusement parks

Columns 206-215: Self-image/self-esteem. Numbers represent how much a user agrees with a statement. Note that if a statement has "reverse polarity", e.g. statement 2 "at times I feel like I am no good at all", it has already been re-coded/inverted by the professor such that higher numbers represent higher self-esteem. Here are the 10 items, in column order:

1 On the whole, I am satisfied with myself
2 At times I think I am no good at all
3 I feel that I have a number of good qualities
4 I am able to do things as well as most other people
5 I feel I do not have much to be proud of
6 I certainly feel useless at times
7 I feel that I'm a person of worth, at least on an equal plane with others
8 I wish I could have more respect for myself
9 All in all, I am inclined to feel that I am a failure
10 I take a positive attitude toward myself

Column 216: User age
Column 217: User gender (1 = male, 2 = female, 3 = non-binary)
Column 218: Political orientation (1 = progressive, 2 = liberal, 3 = moderate, 4 = conservative, 5 = libertarian, 6 = independent)
Column 219: Art education (The higher the number, the more: 0 = none, 3 = years of art education)
Column 220: General sophistication (The higher, the more: 0 = not going to the opera, etc. 3 = doing everything – opera, art galleries, etc.)
Column 221: Being somewhat of an artist myself? (0 = no, 1 = sort of, 2 = yes, I see myself as an artist)

Note that we did most of the data munging and coding for you already but you still need to handle missing data in some way (e.g. row-wise removal, element-wise removal, imputation).
Extreme values might also have to be handled.

Questions management would like you to answer:

1) Is classical art more well liked than modern art?
2) Is there a difference in the preference ratings for modern art vs. non-human (animals and computers) generated art?
3) Do women give higher art preference ratings than men?
4) Is there a difference in the preference ratings of users with some art background (some art education) vs. none?
5) Build a regression model to predict art preference ratings from energy ratings only. Make sure to use cross-validation methods to avoid overfitting and characterize how well your model predicts art preference ratings.
6) Build a regression model to predict art preference ratings from energy ratings and demographic information. Make sure to use cross-validation methods to avoid overfitting and comment on how well your model predicts relative to the "energy ratings only" model.
7) Considering the 2D space of average preference ratings vs. average energy rating (that contains the 91 art pieces as elements), how many clusters can you – algorithmically - identify in this space? Make sure to comment on the identity of the clusters – do they correspond to particular types of art?
8) Considering only the first principal component of the self-image ratings as inputs to a regression model – how well can you predict art preference ratings from that factor alone?
9) Consider the first 3 principal components of the "dark personality" traits – use these as inputs to a regression model to predict art preference ratings. Which of these components significantly predict art preference ratings? Comment on the likely identity of these factors (e.g. narcissism, manipulativeness, callousness, etc.).
10) Can you determine the political orientation of the users (to simplify things and avoid gross class imbalance issues, you can consider just 2 classes: "left" (progressive & liberal) vs. "non-left" (everyone else)) from all the other information available, using any classification model of your choice? Make sure to comment on the classification quality of this model.

Extra credit: Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions.

Hints:

*Beware of off-by-one errors. This document and the csv data files index from 1, but Python indexes from 0. Make sure to keep track of this.
*In order to answer some of these questions, you might have to apply a dimension reduction method first. For instance, "dark personality traits" and "self-image" are characterized by 10-12 variables each. Similarly, you might have to reduce variables to their summary statistics.
*In order to do some analyses, you will have to clean the data first, either by removing or imputing missing data (either is fine, but explain and justify what you did)
*If you encounter skewed data, you might want to transform the data first, e.g. by z-scoring
*To clarify: When talking about "principal components" above, we mean the transformed data, rotated into the new coordinate system by the PCA.
*Avoid overfitting with cross-validation methods.
*How well your model predicts can be assessed with RMSE or R^2 for regression models, or AUC for classification models.
*You can use conventional choices of alpha (e.g. 0.05) or confidence intervals (e.g. 95%) throughout.