STAT 847: Analysis Assignment 2

DUE: Tuesday, March 28 2023 at 11:59pm EST

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically in .pdf format via Crowdmark.

There are 44 marks in total. You might be surprised by how little work 1 mark represents.

For all questions, these are variables to consider.

Response:

Variate	Description
SCORE	The overall score of the building.

Explanatory:

Variate	Description
ENTRANCE_LOBBY	condition of entrance and/or lobby in a building. 1 being the worst and 5 being the best.
ENTRANCE_DOORS_ WINDOWS	condition of entrance doors and windows in a building. 1-5
SECURITY	condition of security system(s) in a building. 1-5
STAIRWELLS	condition of stairwells in a building. 1-5
INTERNAL_GUARDS_HANDRAILS	condition of internal guards and handrails in a building. 1-5
GARBAGE_BIN_ STORAGE_AREA	condition of garbage bin storage room or outdoor enclosure area. 1-5
INTERIOR_WALL_CEILING_FLOOR	condition of internal walls, ceilings and floors in a building. 1-5
INTERIOR_LIGHTING_LEVELS	condition of internal lighting levels in a building. 1-5
GRAFFITI	severity of graffiti in a building. 1 being significant graffiti and 5 being no graffiti.
EXTERIOR_CLADDING	condition of exterior cladding/bricks/paint, flashing and drain pipes on a building. 1-5
EXTERIOR_GROUNDS	condition of exterior grounds of a building. 1-5
EXTERIOR_WALKWAYS	condition of exterior walkways of a building. 1-5
WATER_PEN_EXT_BLDG_ELEMENTS	condition of water penetration of external elements of a building. 1-5

1. (6 marks) Get the latest dataset on Apartment Building Evaluation by searching for "apartments" and finding the appropriate ID from the opendatatoronto API package.

Show your code for `searching`, and for `getting the dataset` as either a csv or as an R dataset, and a `tibble` of the dataset.

If you have trouble with this question, you can skip it and use the dataset "Apartment Evaluations 2023.csv" in Learn at no penalty to future questions. You can always go back and get the dataset later.

Note that some datasets get updated quite often, including this one, which is updated several times a week. The key (and Learn dataset) use the March 13, 2023 dataset. Small changes between the key and your answers are expected.

If you get an error when running the search like: `Error in loadNamespace...`, look at the `namespace`. Close and reopen R, and then `install.packages` the package named after `namespace`. You may need to do this a few times as `opendatatoronto` needs some very recently updated packages. I ran into this with `cli` and `dplyr`.

2. (10 marks) Clean the dataset so that it only contains the relevant variables, as a `data.frame`, and each variable is correctly identified as numeric. (Hint: `names(dataset)` will help you organize this).

Keep only complete cases. That is, keep only rows that have data for all 14 of these variables. Do this step only after you have isolated the variables and turned them into numeric variables, otherwise you may remove too few or too many cases.

Show a `dim(dataset)` and a `summary(dataset)` to show that the data only contains these 14 variables (1 response, 13 explanatory), as numbers, with no NAs.

3. (6 marks) Use best subsets regression with the BIC criterion, select a model using the listed variables as candidates. Report both the `summary(lm())` of the resulting model and the best subsets.

4. (6 marks) Run a PCA on only the explanatory variables from the last question. Report your code and the head of the individual coordinates. Use the same number of dimensions as you used variables in the last question. (Hint: Look to `?PCA` for guidance on how to leave out response variables as supplementary)

5. (4 marks) Build a linear model of the response variable `SCORE` using the first three PCA dimensions from the previous question, and nothing else. Report the summary(lm(dataset)).

6. (8 marks) Describe briefly two advantages and disadvantages of this PCA-based model over the best subsets model from earlier in this assignment. (There are several correct answers, but only the first two will be marked).

7. (4 marks) The variance inflation factor of an explanatory variable in a model is a function of how collinear that variable is with the over explanatory variables in the model are. The higher the number, the more collinear and the most the variance estimates of the slopes are being inflated by including that variable. We can find the variable inflation factor with `vif(lm())`, where `vif` is found in the `car` package.

Find the `vif()` of both the PCA-based model and best-subsets model.

Report the VIFs for both models and briefly why the PCA-based model has such low inflation factors (1 is the lowest possible).