Business Problem

The Orange River Public Library has petitioned the city council of Orange City[1](#_bookmark0) for more funding. The library’s reasoning is written down in the following letter to the city council.

*City Council Members,*

*You have seen that we have requested additional money for several expenses. With increased funding, the library can offer more programs, improve the collection, and hire more staff, which will in turn draw more visitors to the library each year. This will improve Orange City culture with a stronger connection to the library, the community, and increase our community’s love for books. The following list details how we will use the funding:*

*Increase full time equivalent number of librarians from 2 to 4*

*Increase full time equivalent of other paid staff from 1 to 2*

*Increase the book volume from 16,000 to 20,000*

*Keep hours open constant at 1750 per year*

*Increase the number of kids programs from 150 to 250*

*Increase the number of young adult programs from 50 to 100*

*Increase the number of other programs from 100 to 150*

*Instigate an interlibrary loan program where we currently have 0 loans to and from but anticipate 20 loans per year to and from when we start*

*Thank you for considering this funding increase. Sincerely, Head Librarian*

The city council is willing to meet the library’s request. They believe in the same ideals about increased library visitations improving the community. However, they are not so sure about what areas to focus on to increase visitations. For example, is a larger collection typically associated with more visitors or is providing additional library programs more important? For this purpose, they have come to your consulting company.

The city council has come up with 3 specific questions they want answered:

What factors lead to more visits per 100 people in a library’s legal service area such as staff, programs, collection, etc? A legal service area is the region containing the population designated for a specific library.

Where is money typically spent for libraries like Orange River Public Library? How much is spent on staff wages, how much on programs, and how much to build and maintain a collection?

If the changes are made as outlined in the memo from the library, what would the effect be on predicted expenses? The city council wants an estimate but also a range of possible values for expenses should these changes occur.

1 While there are cities named Orange City, this one is fictitious, created for this assessment.

The Institute of Museum and Library Services conducts a yearly Public Library Survey that is publicly available. In this data set, museums provide information on the exact variables the city council is concerned with learning more about, as well as several others. Some of these variables describe the type of region the library resides in. In the context of the survey, Orange River Public Library has legal basis code CI, referring to the library belonging to a municipal government and geocode CI1. It is a single outlet administrative entity for purposes of administrative structure and. It does match the FSCS public library definition. There are about 15,000 individuals in the legal service area for the library.

This data can be found in. the file libraries.csv.[2](#_bookmark1) There are many codes in the data, and so it is important to consult the data dictionary provided in the file library\_data\_dictionary.xlsx.

As is common with survey data, there are a lot of missing values. These are typically missing for three different reasons.

The library was closed during the period being measured.

The information is being suppressed to protect individual information. For example, if a library has one employee, the salary information for that library can be sensitive.

The data point wasn’t provided by the library for an unknown reason.

Use this data to provide answers to the city council. As you perform these tasks write sections of a technical report as described in the tasks. In addition to the sections of the technical report, you will write a high -evel overview for the city council that is less technical and more accessible for those not familiar with the specific language used in predictive analytics.

File List

One .csv file called libraries.csv: Public Library Survey

One .xlsx file called Library\_Data\_Dictionary.xlsx: the data dictionary

2 The data is available at [Public Libraries Survey | Institute of Museum and Library Services (imls.gov).](https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey) The data set provided for this assessment has been modified from the original and is the only one that should be used.

**Task 1**

The data set in its current form is not ready to address the business problem. Specifically. fix the following issues:

* Identify key variables that should be included in analyses to respond to the business problems. For these variables, justify and implement a solution to deal with the missing values. Consult the data dictionary to determine if there is a special kind of missingness for the variables.
* Create the target variable of visits per 100 individuals in the legal service area. Use visualizations to determine if any transformations should be made or to identify any outliers. Justify any transformation or any action taken to address outliers.
* Many of the other continuous variables can also be turned into rates per 100, such as employee numbers, collection size, and the number of programs. Create rates per 100 versions of these variables. Use visualizations both univariate and against the target variable to justify possible transformations of these variables.

Perform a general cleaning and data validation for the data set. Write the technical data preparation section of your report. Because none of your code will be available to the reader, all evidence of the data preparation tasks must be contained in your report. This may include written descriptions as well as charts, however the work is most effectively conveyed.

Task 2

Address the issue of target leakage in the data. Are there variables that should not be used for prediction because they cannot be measured prior to the target variable being measured? Perhaps these are variables that are observed along with visits as opposed to numbers that can be planned in ahead. Be sure to not include those variables in the predictive models you will be building.

Task 3

The first business problem is to determine what factors are most associated with library visits. Address this problem in the following two ways.

1. Build a linear regression or generalized linear model and tune the model appropriately. Use *p*- values and test statistics to make conclusions as to which variables are important in the model.
2. Build a random forest model. Use feature importance measures and partial dependence plots to show the effect of the factors on library visits. Interpret the partial dependence plots appropriately.

Determine where the conclusions of the two models agree or disagree. Write a section of your technical report where you describe the models used and use the model output to answer the first business problem.

Task 4

The second business problem is to determine spending distributions for libraries similar to the Orange River Public Library. Instead of using a predictive model to answer this question, aggregate and summarize the data in useful ways. Filter the data in such ways as create a small subset that is appropriate for answering the question and provide a table and summary that will address the question. Use only libraries where the population in the legal service area is within 2,000 individuals for that of Orange River County Library’s legal service area.

Write a section of a technical report where these conclusions are given and justified.

Task 5

The third business problem is to provide a range of predictions for expenses of the library based on the new features given. Build a Bayesian linear model to address this final question. The information provided in the business problem does not completely match the data given, so when building your model, be sure to use only variables that can be used for the prediction. If you use priors that are not the default priors, justify your choice of priors.

Write a section of the technical report that describes the model being used and provides and interprets the output that answers the business problem.

[Note: The modules provided for preparation for this exam use the BRMS packages and functions for Bayesian models. The BRMS functions work best when the data you use to build and predict has been standardized, meaning the continuous variables have a mean of 0 and a variance of 1.]

Task 6

In no more than three paragraphs, write an executive summary to the Orange River city council that gives a high-level overview of your findings. Include your confidence in the model output and any tables or figures that directly answer a business problem.