Course Project - Airport Restaurant Data Dictionary and Data Governance

Shaun Pritchard

Rasmussen College

QMB4000

Benjamin Tasker

November 15, 2020

**Course Project - Airport Restaurant Data Dictionary and Data Governance**

Below is a description of a data dictionary and report outline the principles and plausible courses of action in implementing a Data Governance strategy for FLIGHTZ company. The data dictionary is a centralized repository of information about data such as meaning, relationships to other data, origin, usage, and format.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset** | **Fieldname** | **Datatype** | **Length** | **Exp.Values** | **Description** |
| WEBEXTRACT | CAMIS | Varchar | 10 | Identifier | This is a unique identifier for the entity (restaurant) |
| WEBEXTRACT | DBA | Varchar | 255 | DBA name | This field represents the name (doing business as) of the entity (restaurant) |
| WEBEXTRACT | BUILDING | Varchar | 3 | Airport building name | AIRPORT Abbreviation links to location. |
| WEBEXTRACT | CUSINE DESCRIPTION | Varchar | 200 | Building number | This field represents the building number for the entity within the airport (restaurant) |
| WEBEXTRACT | INSPECTION DATE | Datetime | 10 | Data of inspection | This field describes the entity (restaurant) cuisine. |
| WEBEXTRACT | ACTION | Varchar | 150 | Actions taken per violation | This field represents the date of inspection. NOTE: Inspection dates of 1/1/1900 mean an establishment has not yet had an inspection |
| WEBEXTRACT | VIOLATION CODE | Varchar | 3 | Violation Code | This field represents the action that is associated with each restaurant inspection. |
| WEBEXTRACT | VIOLATION DESCRIPTION | Varchar | 600 | Description of DBA Violation | This field represents each violation associated with a restaurant inspection. |
| WEBEXTRACT | CRITIAL FLAG | Varchar | 1 | Critical | This field describes the violation codes |
| WEBEXTRACT | SCORE | Varchar | 3 | Overall business score | Critical violations are those most likely to contribute to foodborne illness. |
| WEBEXTRACT | GRADE | Varchar | 1 | Overall business grade | Total score for a particular inspection; updated based on adjudication results. |
| WEBEXTRACT | GRADE DATE | Datetime | 10 | Date grade was given | This field represents the grade associated with this inspection. Grades given during a reopening inspection are derived from the previous re-inspection. |
| WEBEXTRACT | RECORD DATE | Datetime | 10 | Date on record | The date when the grade was issued to the entity (restaurant) |
| WEBEXTRACT | INSECTION TYPE | Varchar | 64 | Type of inspection | The date when the web extract was run to produce this data set |

**Data Dictionary:**

The data dictionary provided compared to the data set provided does have discrepancies. There are missing expected values DBA, BUILDING, CUISINE DESCRIPTION, VIOLATION CODE, SCORE, GRADE DATE, and RECORD DATE. There are null values in the length column which specifies the number of characters and memory allocation space required for a categories field. Those fields denoted by N/A are not valid character settings. The row labels of INSPECTION DATE, GRADE DATE, AND RECORD DATE. These fields are dates that can be formatted using up to 8 bytes in the format of YYYY-MM-DD HH:MM:SS which would facilitate 10 to 12 bits of space using the DATETIME data type. Also, some recommendations of the datatypes being used for the column CAMIS would be to switch the datatype to INT as opposed to VARCHAR. The data set only provides numerical data throughout the entire column without floating decimal points.

**Data Set:**

In the data set itself, we also find missing null values and discrepancies in the column data that will need to be normalized and preprocessed for accurate results. The building column could be numerical except that there are character values, special characters, and numbers used for building codes that do not seem uniform with the building code data in only several occurrences. There are also several occurrences of 0 leading as building number and missing fields of data for this column. The Action column specifies redundant data stating “Violations were cited in the following area(s).” in 99% of all the column data. There are also missing values and empty rows of data. The SCORE column should be of type INT. The GRADE column has many missing values. The GRADE DATE column has many missing values null values as well. Also, the INSPECTION TYPE column appears to have categorized data with defined data types that would need to be hot encoded to run inferences on.

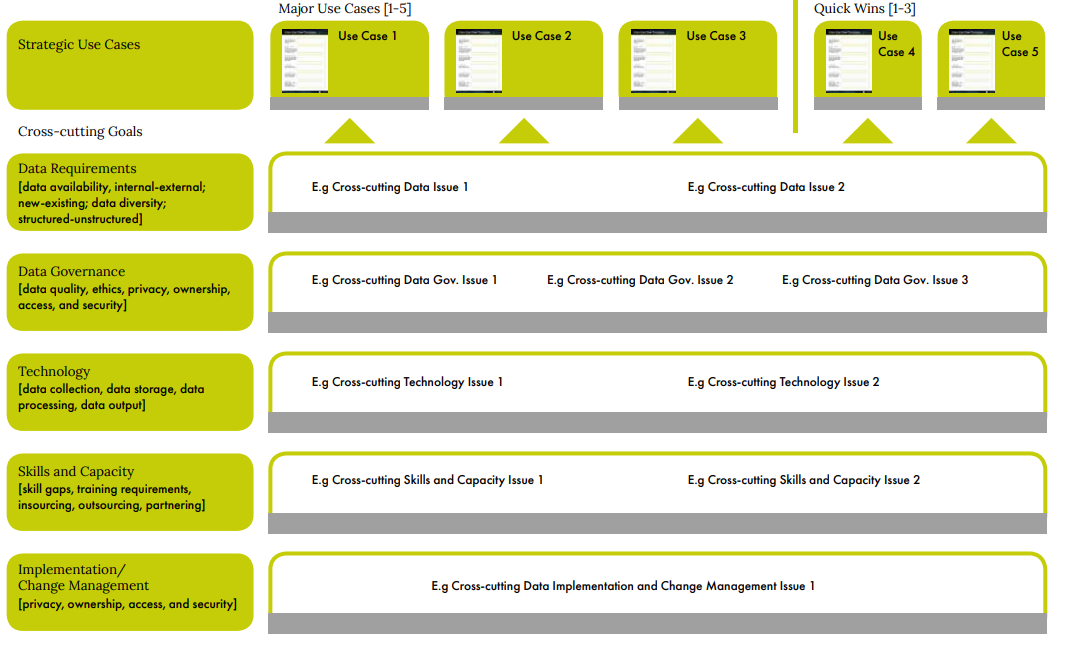
**Data Governance:**

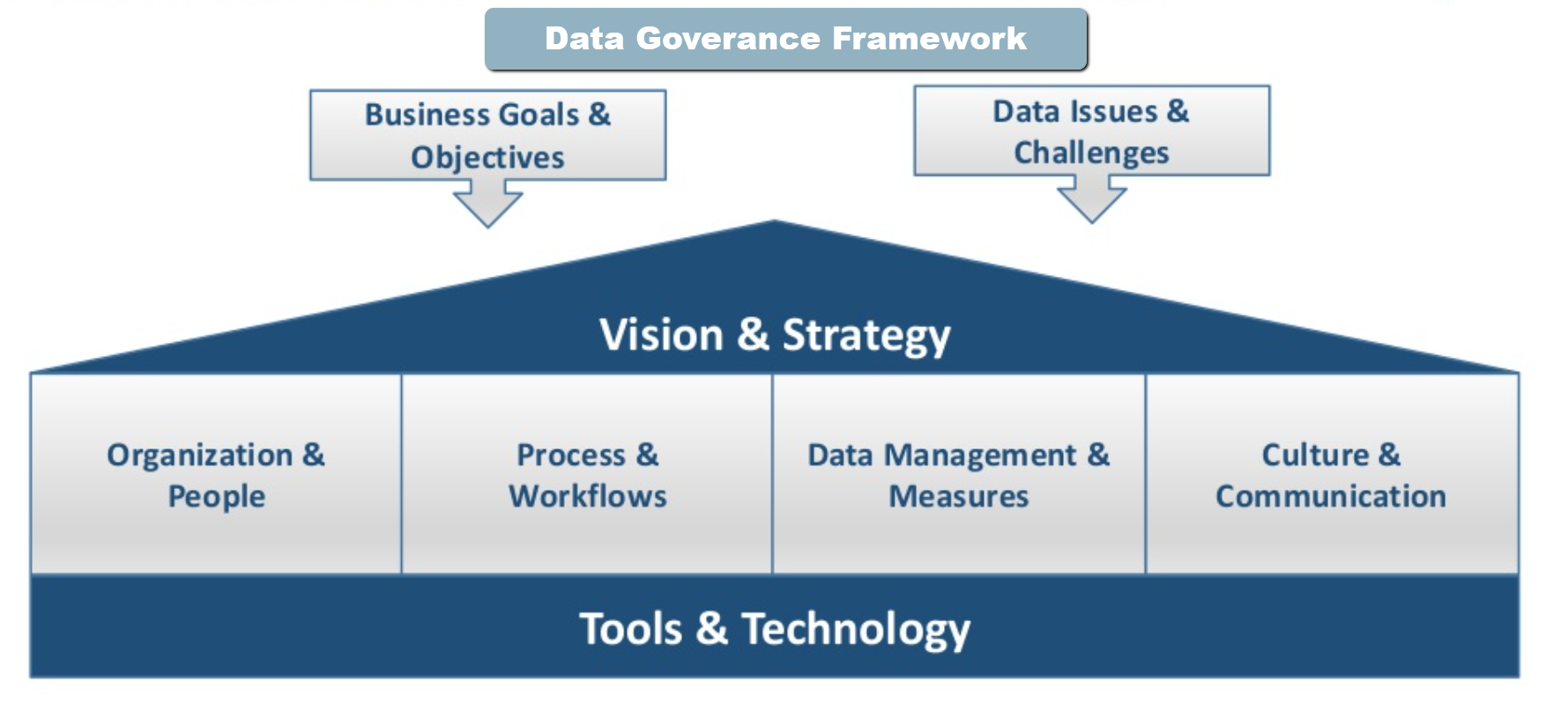
Based on the following data I will give my recommendations to build a data governance strategy and Answer key questions vital for data business infrastructure. First, we need to normalize and clean the dataset to build accurate models and inferences. we will then develop a governance plan to identify the overall strategic goal and hierarchy to move forward. Keep in mind data governance is a business-led continuous process of improving data. The overall goal of Data governance is ensuring that data is accessible, manageable, trusted, and protected for those who need not make critical business decisions.

**Here are the 4 main steps we will need to implement:**

* **Business Goals and Strategy:** Identifying business goals and objectives aligned to data
* **Current State Assessment:** Understanding the current maturity and environment.
* **Proposed Future State:** Propose future state capabilities, processes, and Organizational Structure.
* **Implementation Roadmap:** Prioritizing efforts and identifying “quick wins.”

To implement this, we can use multiple strategies including a template that will define the overall data governance strategy *(Seiner, 2008).*



We will then lead to the development of business capability models. Use data architecture to guide governance. Identifying dependencies with process models. Developing customer Journey Maps in applying a structured data governance framework

A well-designed data governance program typically includes a governance team, a steering committee that acts as the governing body, and a group of data stewards. They work together to create the standards and policies for governing data, as well as implementation and enforcement procedures that are primarily carried out by the data stewards. Executives and other representatives from organizational business operations take part, in addition to the IT and data management teams *(Amy Filener, 2019).*

Data governance is both a technical and organizational discipline that requires a combination of data management fundamentals with organizational change. We can ensure data-governance is successful by

* Having a Clear Vision of what good looks like
* Setting realistic expectations
* Allied efforts with business benefits
* Focus on what size is the data that matters
* Equip people within our company to succeed
* Learning from past practices
* Recognizing that there is no one-size-fits-all
* Used recognized industry Frameworks

Through this implementation, these are the recommended solutions, models, and recommendations given the current data and expressed goals from FLIGHTZ company.

**Fixing NULL values, missing values:**

There are many techniques we can use to implement the missing and no values within the data. We have two types of data that would apply to our inferences numerical data and categorical data. And then we have descriptive data.

I propose using python libraries (Pandas, NumPy, Scikit learn) to model the data to vectors and complete analysis that will yield important business information. Here are several of the methods we can use to fix this issue *(Grace-Martin, n.d).*

* **Do Nothing** - some data will not be relevant in making business decisions for this data we can just simply not use it.
* Creating dummy variables - dummy variables works by associating a numerical variable to represent categorical data. for example, building columns of data to represent states California and New York we could represent California with 01 and represent New York with 10
* **Imputation Using (Mean/Median) Values** - by calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. It can only be used with numeric data.
* Imputation Using (Most Frequent) or (Zero/Constant) Values - Most Frequent is another statistical strategy to impute missing values with categorical data. by replacing missing data with the most frequent values within each column.
* **Imputation Using k-NN** - The k nearest neighbors is an algorithm that is used for simple classification. The algorithm uses ‘feature similarity’ to predict the values of any new data points.
* **Imputation Using Multivariate Imputation by Chained Equation (MICE)** - This type of imputation works by filling the missing data multiple times. Multiple Imputations (MIs) are much better than a single imputation as it measures the uncertainty of the missing values in a better way.
* **Imputation Using Deep Learning (Datawig)** - This method works very well with categorical and non-numerical features. It is a library that learns Machine Learning models using Deep Neural Networks to impute missing values in a data frame.

Using python library methods such as scikit learn to use built-in imputation functions to complete missing and null data. Scikit learns has powerful computational methods to encode blanks, NaNs, placeholders of inconsistent data sets with real estimators which assume values in the column array of values. More precise variances using imputation calculations *(Scikit-Learn, 2020)*.

**Some of the methods available in this library are:**

* One Hot Encoding
* Multivariate feature imputation
* Univariate feature imputation
* Single Imputation
* Nearest neighbors’ imputation
* Marking imputed values

**Training Employees:**

It is essential to onboard and train employees with higher education of data governance practices. Data governance is an institutional priority and it is critical to give all employees the resources and training to achieve this priority as soon as they walk in the door *(Varonis.com, 2020).*

**The data governance initial training can be broken down into 4 parts:**

* **First Training** – If possible, have a face to face meeting (or phone call) with the employees
* **Assignments** – At the first training, provide the new employee with a list of assignments that they can do when they have some spare moments.
* **Follow-Up Training** – Maybe after a prolonged period schedule have another face to face meeting (or a phone call) with the employees to see if they have any data governance-related questions.
* **Resource Page** – Publish a web page or internal document (wiki) that contains information that all employees can access training, data governance practices, priority models, and other established data governance policies.

**Important Metrics FLIGHTZ should use:**

FLIGHTZ initial goal is to learn more about the restaurants at the airports and the health inspection violations. We need to define a strategy to which end this data will be facilitated. For this, we need to ask some more questions. What exactly are we looking for concerning violation? (repeat offenders, quality of food, etc.) What issues can we solve with this data, and what goals does the CEO want to meet.

There are many underlying analyses that we can reach with this data if there was more definition involved from the CEO. Here are my recommendations on the metrics that we should be focusing on to achieve these types of results. I will assume this analysis identifies issues, redundancy, recurrence, and quality.

I think we should find correlations using linear regression between the inspection dates and the business scores for numerical data and the business name for categorical data to tell us which restaurants are getting more violations.

We could use metrics based on the grades and grade dates to find the frequency of times more issues occur across any business serving food. We can also determine violation frequency for specific repeat businesses.

We can use multiple linear regression to combine more samples between numerical and categorical data to run multiple inferences or determine truth based inferenaces and predictions based on multiple linear regression models and performance factors.

We can encode data to represent values such as the critical Flags, inspection types, actions, and even the cuisine descriptions. Through this, we can determine which businesses are getting more violations, which types of cuisines are getting more violations, which locations are getting more violations to determine which businesses in specific locations are meeting standards of inspection. Which business is doing better and excelling they stopped their location their type of Cuisine and the record track have inspections. We could run a hypothesis test to determine the variance between inspection types per location, business, and cuisine.

We can use a grading scale to determine which businesses are most successful overall and during specific times of the year. giving us analytics and data supporting to invest more in those locations. Overall, we can determine if there are specific businesses that are not performing does pacific standard but maybe costume money overall using decision tree regression models. We can then determine correlations that could generate forecast data and focus on preparing failed violation businesses that I have the means to be very productive in specific locations during specific times. Also, use the data for modeling future data and reports to set linear discriminant analysis models, machine learning models, and clustering models.

**How to reach successful data competition:**

To reach success in the competition, we would use this data to understand score predictions forecasting & analysis to determine certain key factors that give the company-specific advantages or disadvantages correlating to the given data sets.

Through this, we can determine and forecast different times within a year which specific types of businesses. Based on performance can better enhance their products and capabilities in the market and accelerate those businesses. It allows us to look at underperforming businesses and cut losses. Essentially, giving us more room inability to facilitate hosting successful business. That serves more popular and productive cuisines.

The data can show us which key factors are most frequent within specific airports and specific businesses. Emulation to the types of inspection codes grades in violations giving us an upper hand to turn those businesses which are more profitable. Using this data, we can set and train models that we can use for future data and comparative analysis and have the upper hand and predictive analytics for the FLIGHTZ company in general.

# References

Amy Filener. (2019, Apr 29). *DAS Slides: Data Governance and Data Architecture – Alignment and Synergies*. Retrieved from https://www.dataversity.net/: https://www.dataversity.net/das-slides-data-governance-and-data-architecture-alignment-and-synergies/

Grace-Martin, K. (n.d). *Seven Ways to Make up Data: Common Methods to Imputing Missing Data*. Retrieved from https://www.theanalysisfactor.com: https://www.theanalysisfactor.com/seven-ways-to-make-up-data-common-methods-to-imputing-missing-data/

Scikit-Learn. (2020). *Imputation of missing values*. Retrieved from https://scikit-learn.org: https://scikit-learn.org/stable/modules/impute.html#marking-imputed-values

Seiner, R. (2008, Sept 29). *The DGI Data Governance Framework*. Retrieved from https://tdan.com/: https://tdan.com/the-dgi-data-governance-framework/8688

Varonis.com. (2020, March 29). *Data Governance Framework Best Practices, Definitions and Examples*. Retrieved from https://www.varonis.com: https://www.varonis.com/blog/data-governance/