1. DATA OVERVIEW

There are two datasets: red and white variants of the Portuguese "Vinho Verde" wine. For the purposes of analysis and data transformation red wine dataset was chosen for further processing.

Dataset consists of 1599 instances and 12 variables, as shown in below example:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
| 1. | 7,4 | 0,70 | 0,00 | 1,9 | 0,076 | 11 | 34 | 0,9978 | 3,51 | 0,56 | 9,4 | 5 |
| 2. | 7,8 | 0,88 | 0,00 | 2,6 | 0,098 | 25 | 67 | 0,9968 | 3,2 | 0,68 | 9,8 | 5 |
| 3. | 7,8 | 0,76 | 0,04 | 2,3 | 0,092 | 15 | 54 | 0,997 | 3,26 | 0,65 | 9,8 | 5 |
| 4. | 11,2 | 0,28 | 0,56 | 1,9 | 0,075 | 17 | 60 | 0,998 | 3,16 | 0,58 | 9,8 | 6 |
| 5. | 7,4 | 0,70 | 0,00 | 1,9 | 0,076 | 11 | 34 | 0,9978 | 3,51 | 0,56 | 9,4 | 5 |

Table 1.1 Wine dataset sample

First eleven variables are a continuous data and the quality variable is a categorical column with bell-shape distribution. Most values are in 5th and 6th class:

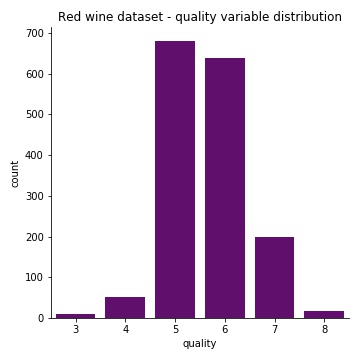


Figure 1.1 Quality variable distribution

2. DATA GENERATOR

In order to store more varied instances of data on Azure Data Lake Gen2 Blob Storage a data generator, in Python programming language, was created. At first it deletes quality variable, next chooses randomly (with repetition) 5000 rows of original red wine dataset, and creates additional variables for each instance:

* datetime (20190101\_000000 to 20191231\_235959 range),
* an arbitrary location (from 0 to 9),

which are then stored as a file name, for example: “2019011\_160512\_7.json”.

Next each of the original eleven continuous variables is randomly modified, stored in JSON format and sent to “rawsink” storage container.



Figure 2.1 Red wine metrics data generator

3. MACHINE LEARNING MODELING

A following step in the process of constructing a modern data pipeline is creating a machine learning model which can later be used to automatically label the wine quality from incoming JSON files.

A Random forest classifier was chosen as a method able to perform multiclass classification and originally implemented in Apache Spark MLlib - scalable machine learning library. Random forests combine many decision trees to reduce the risk of overfitting [5].

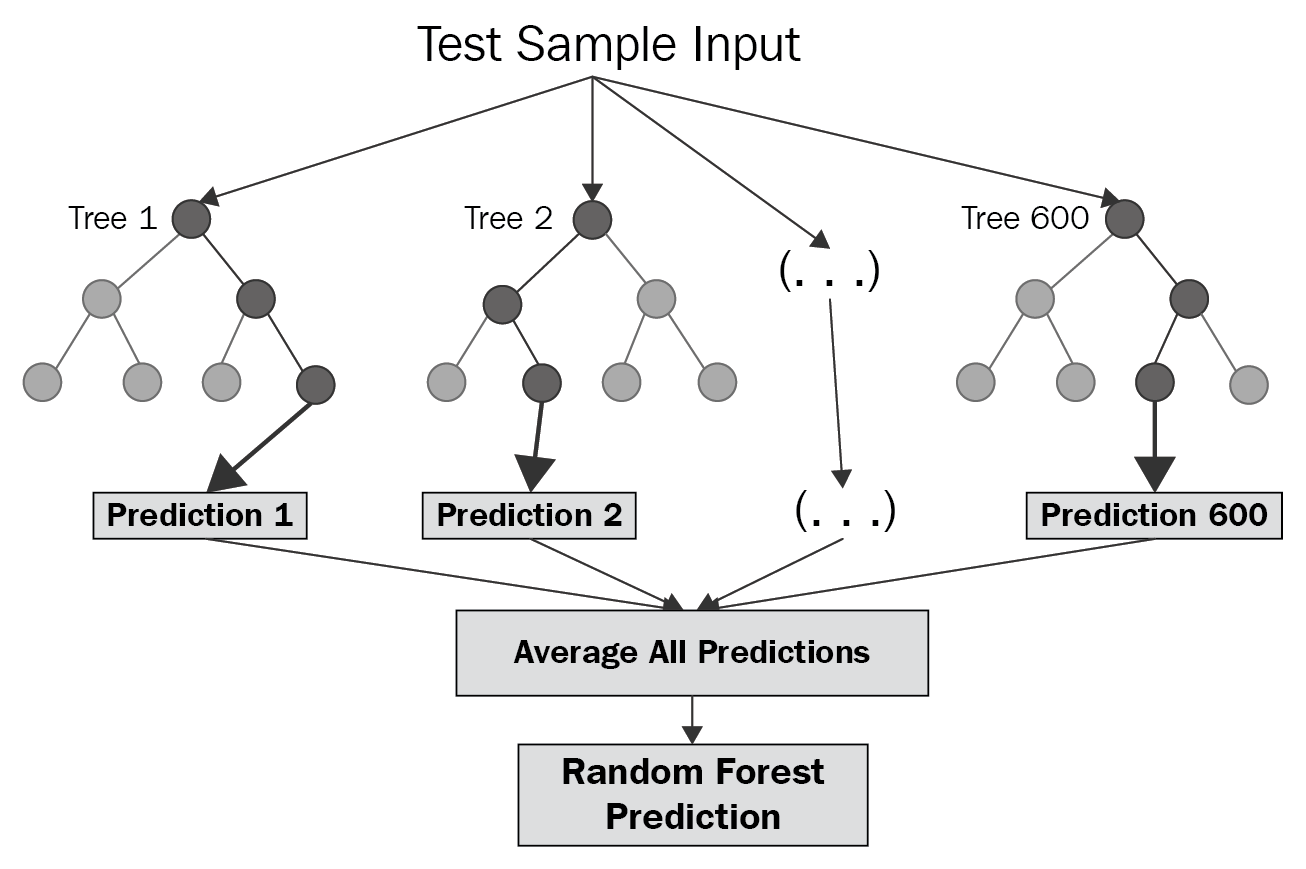


Figure 3.1 Random forest classifier schema [6]

Because fine tuning machine learning predictive model is a crucial step to improve accuracy of the forecasted results, original wine dataset was first split into train / test data (80/20 ratio) and with the usage of grid search on four hyperparameters (impurity, number of trees, maximum tree depth and maximum bins parameters) was next was fit to create the best possible output. All machine learning processes were created with Apache PySpark - the Python API for Spark.

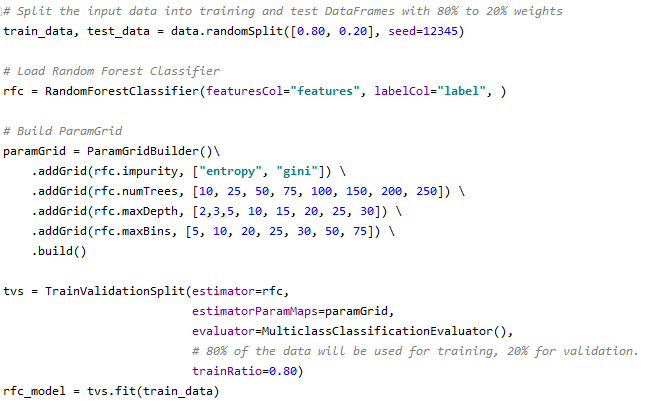


Figure 3.2 Random forest classifier with hyperparameters tuning

The final Random forest classifier has an accuracy: 0.6405 and RMSE: 0.6803, which is only calculated for multiclassification problem to illustrate that the model, when wrong, is usually one class below / above the original quality label, as shown in below chart:

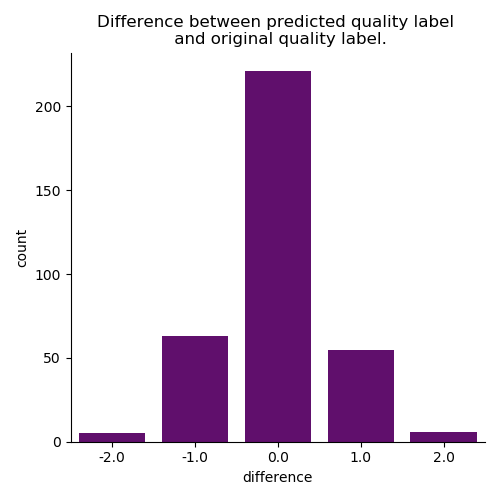


Figure 3.3. Difference between predicted quality label and original quality label

The best model was constructed with below hyperparameters:

* impurity: gini,
* number of trees: 75,
* maximum tree depth: 15,
* maximum bins parameters: 30.

5 - <https://spark.apache.org/docs/2.2.0/ml-classification-regression.html#random-forests>

6 - <https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f>

4. DATA PIPELINE WITH MICROSOFT AZURE DATA FACTORY

With raw JSON files stored by now on Azure Storage Account and ready to use machine learning model for data evaluation, the next step is to build data pipeline for automatization of the whole process. As a case study we can image that raw readings of data parameters come each day to our rawsink Blob Storage from which they should be next transformed, evaluated and pushed to more readable format, available for SQL queries, for more complex data exploration.

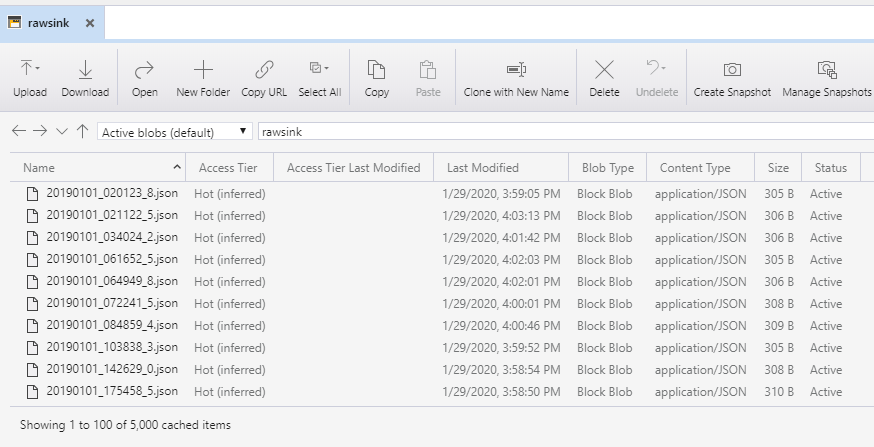


Figure 4.1 Raw JSON files stored on Azure Data Lake Gen2 Blob Storage

For this purpose, we will use **Azure Data Factory** – a cloud-based ETL and data integration service that allows you to create data-driven workflows for orchestrating data movement and transforming data at scale [7]. The core schema of the data pipeline was adapted from “Integrating Data in Microsoft Azure” course by Marcelo Pastorino [8].

First, we create **Get Metadata** activity which connects to our rawsink storage and lists all the items (files and folders) in our storage. Next, just in case any folder schema is sent to the rawsink, we use **Filter** activity to filter out folders from the items returned by the Get Metadata activity. As we cannot delete all files at once from our rawsink container, because it can receive new files while we are running the process, we will move files between two storage containers rawsink and sinkstage to guarantee that each file is processed only once. A **ForEach** activity is used for this process.

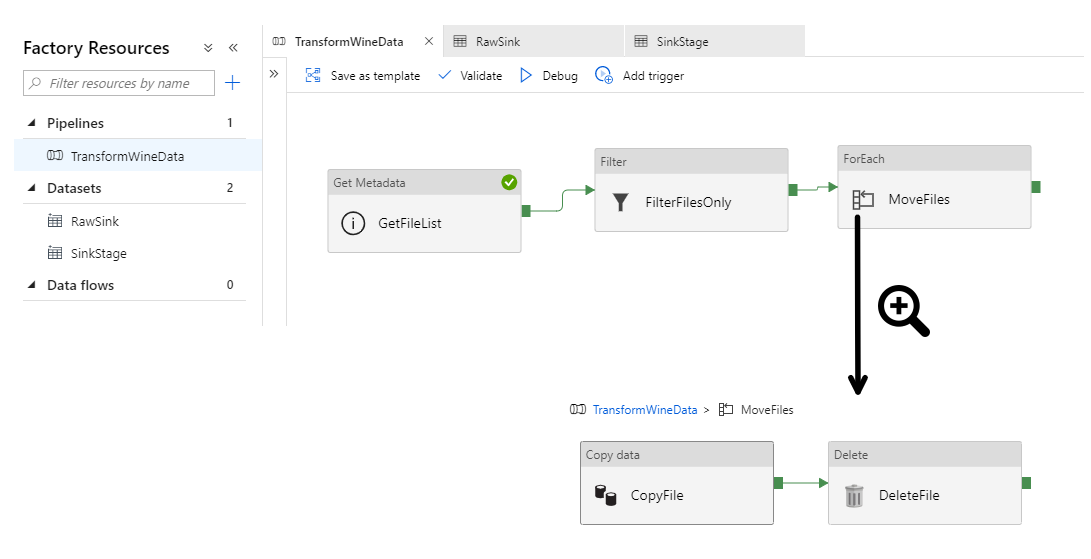


Figure 4.2. Moving JSON files between containers in Azure Data Factory pipeline

After the filtered files were move to proper storage container, advanced data transformations and wine data quality evaluation will be done inside **Azure Databricks** notebook, which will be connected to previously created pipeline. Azure Databricks is an Apache Spark-based analytics platform optimized for the Microsoft Azure cloud services platform to provide one-click setup, streamlined workflows, and an interactive workspace [9].

Initially, a **Databricks cluster** - a set of computation resources and configurations needs to be built in order to run data engineering, data science, and data analytics workloads. For the purpose of the thesis a simple cluster with one worker was created.

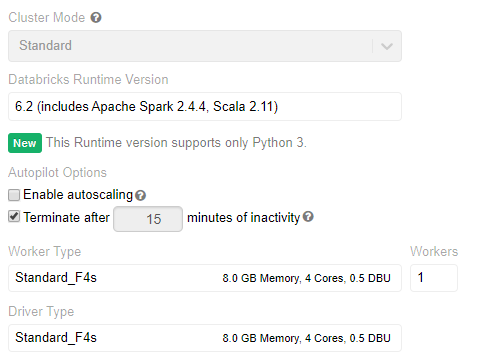


Figure 4.3 Databricks cluster configuration

Next, a new notebook with the code to load, transform and store data back into output database is created. First step is to load all the JSON files to single Data Frame object and enrich the dataset with date, time and location columns. Following, locations are being mapped with corresponding values from provided dictionary.



Figure 4.4. Raw JSON files loaded and transformed in Databricks notebook

Next a VectorAssembler of wine parameters and earlier created Random forest classifier are prepared to predict wine quality.

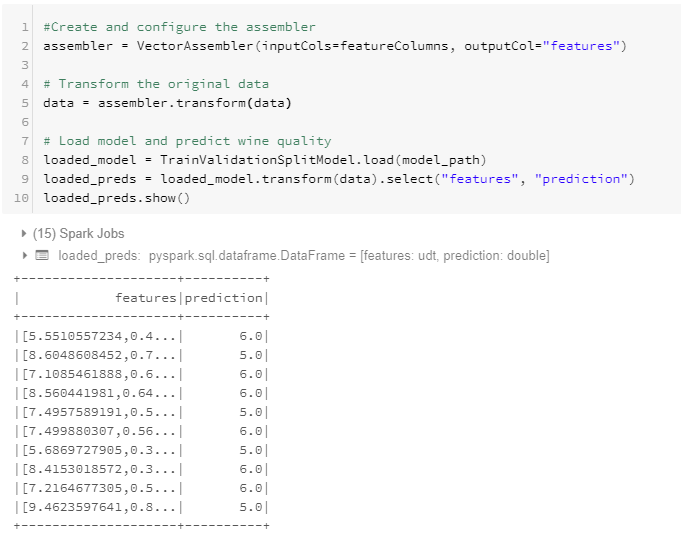


Figure 4.5 Wine quality predictions

Lastly new Data Frame is being created to store sorted and evaluated data.



Figure 4.6 Final Data Frame output

And next it is sent to **Azure SQL Database** - a general-purpose relational database, provided as a managed service, which enables to process both relational data and non-relational structures, such as graphs, JSON, spatial, and XML [10].

The last step in the Azure Data Factory pipeline, which ensures that the JSON files are processed only once, is adding **Delete** activity to delete process files from the stagesink container. Whole process (presented in appendix number 1 to the thesis) can be tracked in real time using **Monitor** panel.

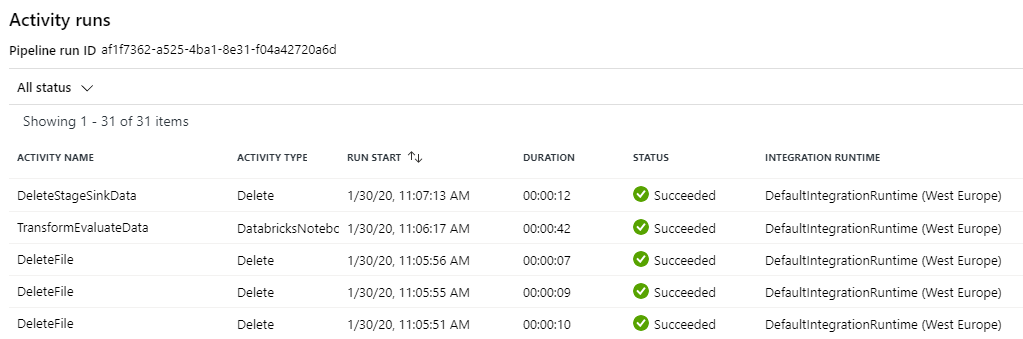


Figure 4.7 Azure Data Factory Monitor panel

7 - <https://docs.microsoft.com/en-us/azure/data-factory/introduction>

8 - <https://app.pluralsight.com/library/courses/microsoft-azure-data-integrating/table-of-contents>

9 - <https://docs.microsoft.com/en-us/azure/azure-databricks/what-is-azure-databricks>

10 - <https://docs.microsoft.com/en-us/azure/sql-database/sql-database-technical-overview>

5. MICROSOFT POWER BI DATA DASHBOARD

Since the input data has been transformed, evaluated and stored on Azure SQL Server it is possible to prepare additional SQL queries and create interactive dashboards.

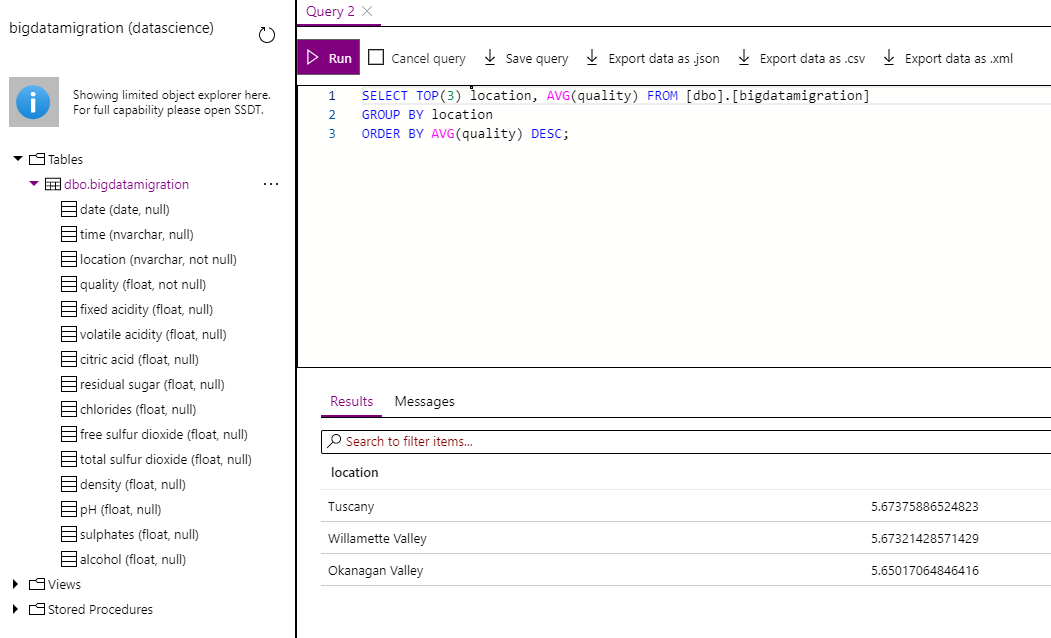


Figure 5.1 Azure SQL Database example query

For graphical exploration it is possible to use **Microsoft Power BI**, connect directly to Azure SQL Database and create reports that use live data. Microsoft Power BI is a business analytics solution that allows good quality data visualizations  and sharing insights on website. For the work objective a simple dashboard with direct access to bigdatamigration database.

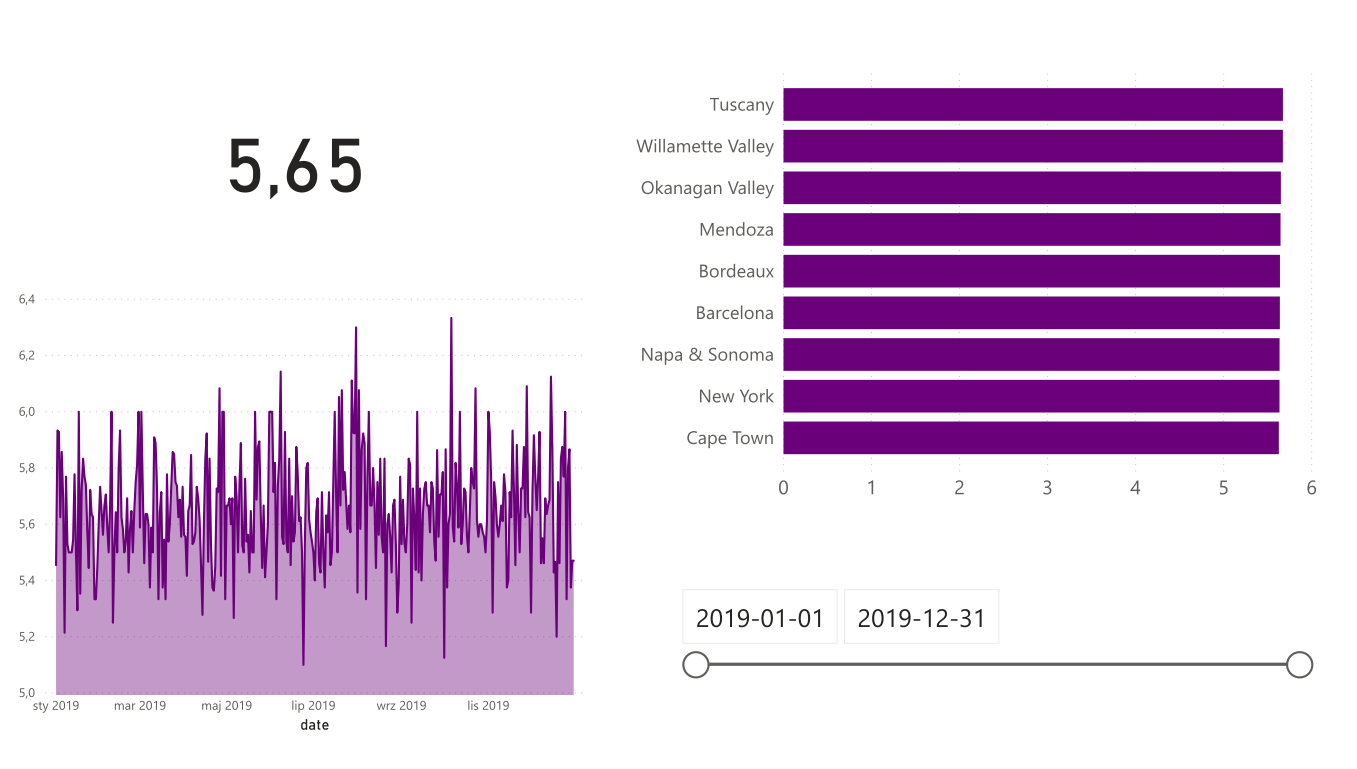


Figure 5.2 Power BI example dashboard

Moving around the interactive panel it is also possible to select a specific date range and filter only the location we want to analyze.

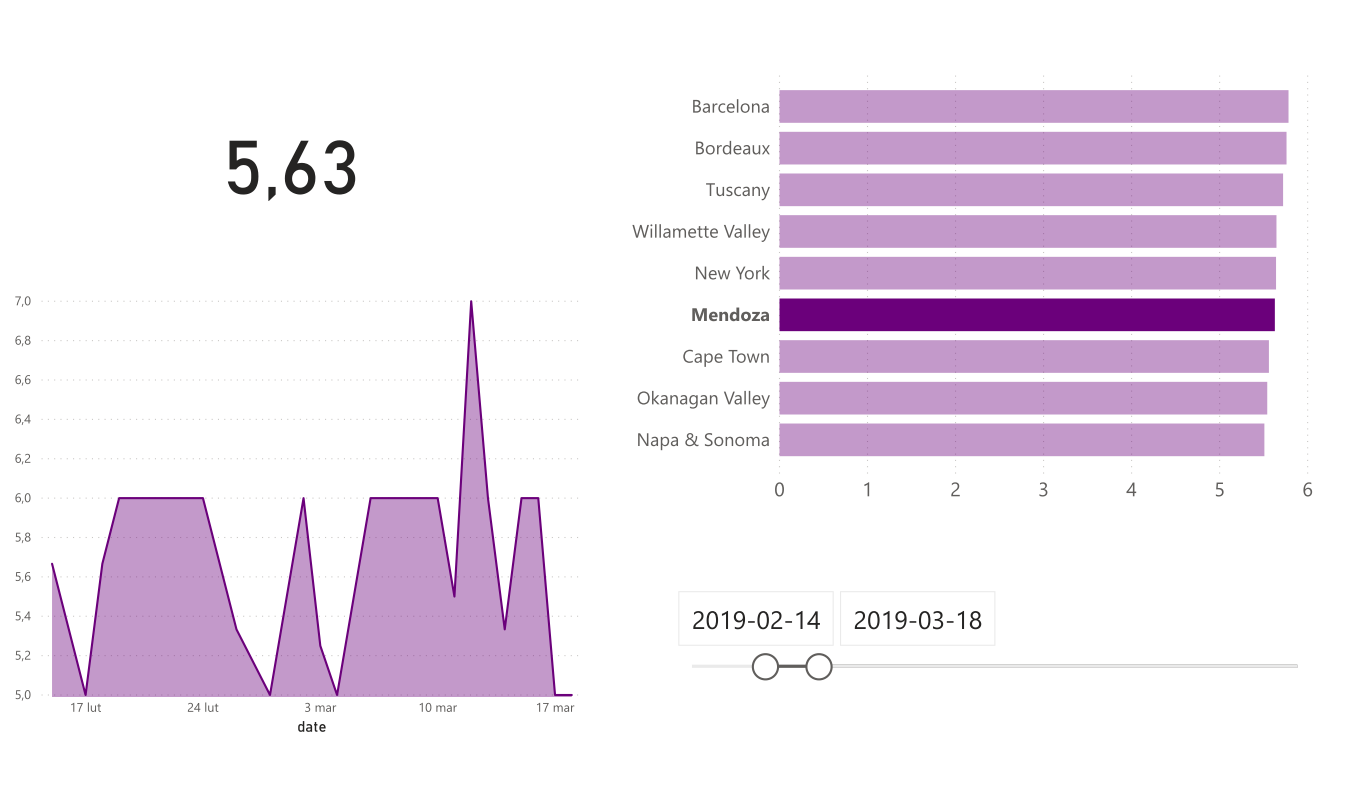


Figure 5.3 Power BI selection example dashboard

Other worth mentioning option is an integration of Power BI with ArcGIS, which offers enhanced mapping and analysis capabilities, demographic data, and compelling visualizations. An example of spatial based dashboard is shown below:

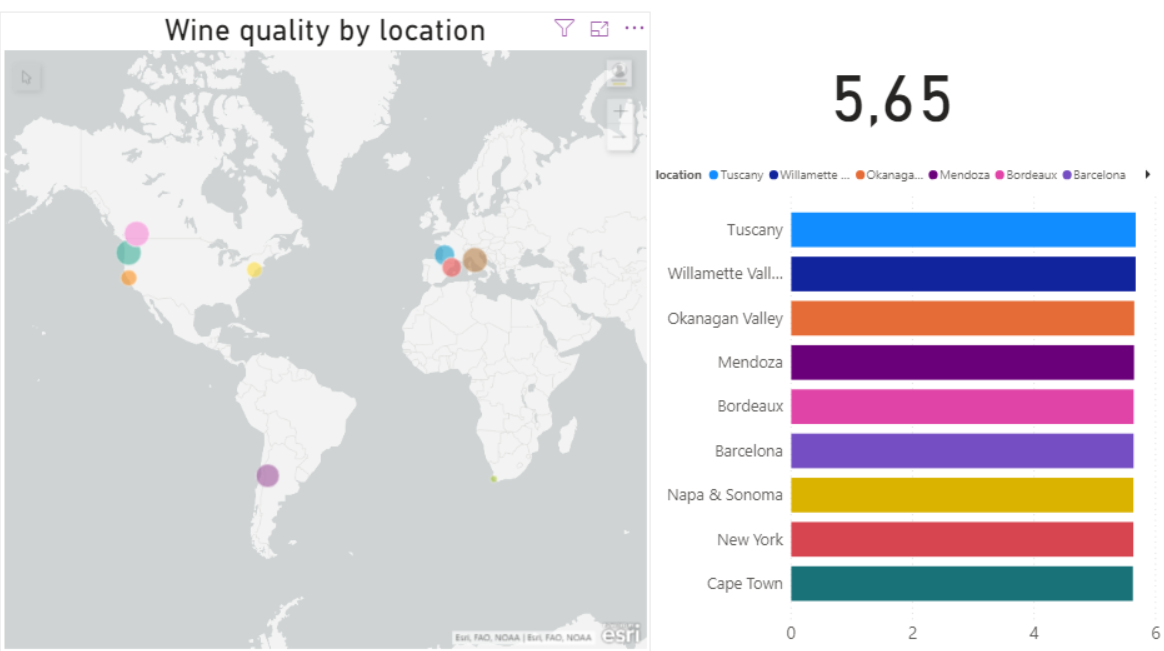


Figure 5.4 Power BI with ArcGIS example dashboard

6. CONCLUSIONS

The phrase "data is the new oil" has become a common refrain and Big data is not just a futuristic terminology anymore. Consequently, we need effective and reliable techniques to get the most value out of our data. Combining technologies like artificial intelligence, machine learning, blockchain, augmented reality, IOT and many more to our classical databases is necessary for a modern enterprise in order to create new products and build an advantage on rapidly evolving market.

In the thesis an approach to combine tools available in Microsoft Azure cloud service was presented, as Microsoft enables simple, scalable, and reliable data pipelines in Azure Data Factory using a serverless, hybrid, parallel data movement service to create new modern ETL workflows in an intuitive working environment. As the final project a fusion of ETL, machine learning and visualization dashboard were shown as output in automating work with data topics.

In summary, combining data engineering and data analysis with AI algorithms in automatized, scaled cloud environment is assuredly one of the most important trends for the near future.