1. File Structure

An attempt has been made to make a design that makes the code easier to read.

There are three Jupyter files: **Public_records_analysis.ipynb**, **Loan_performance.ipynb** analysis and **Merged_tables.ipynb,** in the first two, the respective datasets are analysed and data transformations are performed on them. These transformations are saved in .csv files.

Once the separate analyses have been done, both files are merged into a single file called **Merged_tables.csv** where they are analysed in conjunction.

2. Main Ideas

Throughout the project we have tried to avoid the black box model. An attempt has been made to find a meaning for all the variables and the meaning of the tables has been hypothesized.

The hypothesis made is that the **Public_records.csv** table represents some type of transaction carried out during a bankruptcy process and the **Performance_loans.csv** table represents loans to people who have declared bankruptcy, with the categorical variables being the history of how the loans have been repaid.

The author of this paper does not strongly commit to these hypotheses since, as explained in the files, there are arguments against it (specified in section 3 of **Public_records_analysis.ipynb**) but due to the lack of context it has been decided to work this way. (The author of the text understands that launching hypotheses and checking their plausibility is one of the main objectives of this exercise).

3. Questions

Answering the questions in the email.

3.1. Are there any anomalies in the data? What are they and how did you find them?

Everything and nothing can be an abnormality due to the lack of context and the lack of a clear objective to seek. However, based on the hypotheses made, the main anomalies found would be:

- **Missing primary key in Public Records Analysis table:** It has been found by checking different columns and looking for repeated values.
- **People with too many bankruptcies:** It has been assumed that *analytics_matchkey* refers to an individual. Counting tuples with same *analytics_matchkey* and different *filed_date*

- **Data not correctly encoded:** Categories like *industry_code_kind_of_business* or *reporting_subscriber_code* are not correctly decoded. There has probably been a problem with a low-level language like C.
- **Lack of dates and payments.** It has been assumed that the *Public_Records* table are transactions carried out and, in a transaction, the two most important variables are the amount of money and the date on which it is carried out.

In future lines it is possible to continue searching for anomalies in data in which *analytics_matchkey* matches where changes to files, file source or type of person ('I' or 'C') occur for no reason.

3.2. Pick a variable to investigate and evaluate the distribution - is there anything interesting?

Several variables and combinations of them have been evaluated. As mentioned before, the existence of multiple bankruptcies coincides with their signature date.

More ideas have been:

- As the existence of a Data in Amount correlated well with *public_record_source_type_code* when it takes the value FE and the record type *public_record_type_code* takes the value 7X. Which may mean that certain trials and certain cases publish the Amounts while others do not. Also, the distribution of Amount suggests the existence of errors in the variable.
- That *public_record_type_code_CJ* has a very strong correlation with there being a payment date may mean that this type of document or judgment does not have a payment associated with it or that it is private.
- Verification of correct registration by checking that the dates follow a correct order.
- There is no strong correlation between the existence of data in *Paid_date* and Amount when a priori it should exist.
-

3.3. Are there any relationships between the performance metrics?

The most obvious relationship is that if the payment is delayed for more than 90 days, it must also be delayed for 60 and 30, the same for 60. It has been proven that this occurs for all tuples and it is so.

Late payments do not appear to have a strong explanation for money being lent, with smaller loans being slightly more likely to default. (This makes sense because larger amounts will only be lent to profiles in which there is great security of repayment)

The ANOVA analysis suggests that there is a statistical difference between the means, which makes sense due to the explanation given above.

4. Link between Loan_Performance and Public_Records

As explained, it has been hypothesized that *analytics_matchkey* is a person who has declared bankruptcy and then requested a loan.

The way both tables have been linked has been by taking the last date of the last bankruptcy (it has been assumed that it has only been lent in the last one). As there were still multiple tuples, it was decided to randomly choose which one to choose for lack of a better criterion.

5. Conclusions

i. The lack of a primary key makes it very difficult to confirm a theory of the meaning of tuples in Public_records.
ii. There are relationships between the types of documents and their source.
iii. There may be people with multiple bankruptcies.
iv. Amount can tell us that there is incorrectly collected data.
v. The amount lent gives some explainability regarding the collection of payment (not to be confused with the fact that if more money is lent there is a greater probability of repayment, but rather more money is lent to the person who returns it with more certainty).
vi. Because the simple search for outliers has not been especially fruitful, more techniques should be tried.

6. Future Work

Continuing with the analysis of the data, one could try to do factorial ANOVAS to check the influence of the 'amount' on the type of judgment or the roles of the judgment. Detection of outliers with more sophisticated techniques such as MCA combined with K-means. And detection of points with great leverage with statistical techniques such as Cook's Distance.

Additionally, this project could continue with the proposal of models to predict between the four categories. The author would propose two possibilities.

Multiple logistic regression: with advantages such as the interpretability of its factors, its simplicity and its output in the form of probability. Although it would imply having to do more

work on collinearity and outliers since, like all linear models, it tends to be very sensitive to these problems.

XGBoost: for its great performance in regressions and classifications, usually better than other machine learning algorithms, and for its robustness against problems such as collinearity and outliers. It can also give output in the form of probability. Although with the problem of non-interpretability.

The author of the text would choose multiple logistic regression for its greater interpretability and would use the XGBoost peroformance as benchmarks to adjust the model. Of course, for both models, an argumentation of the data will have to be carried out so that the categories are balanced.