Summary of findings from EDA

**Key findings from EDA:**

* Those earning 60% of their income from music are most likely to be earning mid-level incomes on the £0 - £100,000+ scale, and they do so whilst spending less of their working time on their music than those earning 100% of their income from music. They earn higher mean and median incomes (excluding outliers). They are also the most efficient at earning money (defined as money earned divided by time spent), 42% more efficient than those earning 100% of their income from music and 35% more efficient than the second highest category.

A graph with different colored squares

Description automatically generated with medium confidence

* Those with both primary and secondary music incomes earned higher than those listing no secondary income sources, with means of £15,763 and £10,457 respectively. Both median and median incomes were higher for those with multiple income sources, across every income source category. This suggests that multiple income sources are important for earning higher incomes.
* A collage of graphs showing the different income and the primary income

  Description automatically generated with medium confidence
* Broadcast usage as a higher percentage of income, and being signed to a major record label, were both also associated with higher incomes.
* In terms of negative correlations, self-release/DIY status was negatively correlated with income at -0.2, and inactivity between 2015-2021 negatively correlated with income at -0.15. Band members earned the least of all creator categories, and solo artists the second lowest.

**Summary of the problem for machine-learning**

* The data as it is currently structured does not lend well to linear regression models. Whilst there are patterns in the data as shown in the EDA, these are not linear and it is not clear whether they could lead to a predictive model.
* The wide spread of incomes for any given feature seem to mean that predicting the income for any given respondent is very hard, with models estimating the rough shape of the data but unable to predict higher incomes at all. On the test data, errors between the true and estimated income values increase as the true values increase, with a high correlation between the true value and the error in the predicted value of 0.75.

**EDA in more detail**

There are some features which do have significant correlations with the target variable. The most significant are:

Positive associations

* Percentage of working time spent on music vs mean average income
  + Those spending 60-70% were the most likely to achieve middle earnings, though less likely to earn very highly than than those spending 100%
* Being signed to a major label was the stronger correlation with income out of the easily available numeric columns;
* Broadcasting (e.g. TV and radio use of music) was quite strongly correlated with income. This was borne out in showing that mean incomes were the highest for those who listed this as a primary income source, compared to any other primary source. For those listing this as a secondary income source, it was associated with the middle of the distribution of mean incomes.
* A most surprising finding was that those earning 60% of their overall income from music are performing the best compared to all other percentages, only performing less well than those earning 100% of their income from music on the number of musicians earning the highest incomes. In the highest earning quartiles, those earning 100% of their income from music outperform those earning 60%. So, whilst likelihood of a respondent earning the highest income when spending 60% of working time on music is very low, the likelihood of that respondent earning a middle income is higher than any other percentage category.
* These respondents were also the most efficient at earning money from music, defined as money earned divided by percentage of working time spent. They were 42% more efficient than those earning 100% and 35% more than the next most efficient group (70% of time spent on music).
* A second relevant finding was from primary and secondary income sources. Here I found that those with both primary and secondary incomes outperformed those with only primary and no secondary incomes, at mean incomes of £15,763 and £10,457 respectively. This is therefore a strong predictor of higher and lower incomes.
* In examining social media followers, more followers generally correlated with higher incomes though the lack of respondents with 1,000,000+ followers meant that this category had too few datapoints to be reliable. Median incomes increased for each category up to and including 100,00 – 1,000,000 followers, apart from those with no followers outperforming those with 1-20.
* There was also a strong correlation between number of tracks available on streaming services and income, with a correlation value of 0.36.

Negative associations

* Inactivity in recent years was quite strongly negatively correlated with income at -0.15, along with Self-release\DIY status for releasing music at -0.2.
* A reliance on Live Performances as a source of income in 2019 also was correlated with lower average incomes, though I disregard this due to COVID lockdowns undoubtedly skewing the result.
* Band members earned the least of all creator categories, with solo artists in second lowest. There were some differences between other categories but within a narrow range.

Other notable results

* Length of career showed no clear correlation to average income.