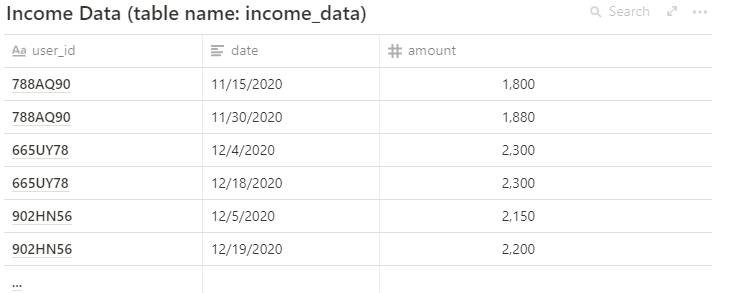
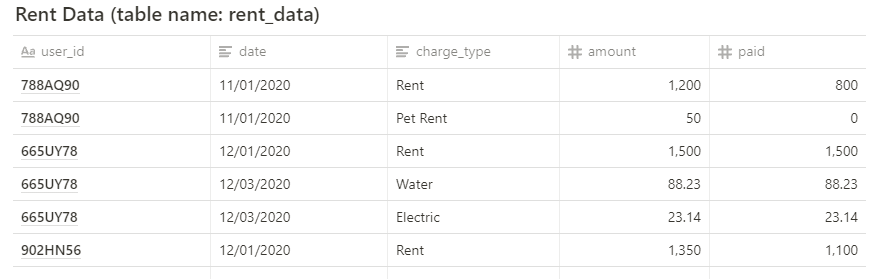
## **Problem #1**

Till has a large data ecosystem of diverse datasets. One of our core value propositions to the marketplace is our ability to take disparate sources of data, get them to talk to one another, and ultimately garner greater insight *because* of one another.

For example, if I have income data over the past 24 months on 100 residents in a building (panel) and the building itself has 300 residents with rent data over the same time period (panel), how would you tackle the following problems below?

Here are the first 6 rows of each dataset:





## **Exercise**

1. Provide the SQL code that generates the rent-to-income ratios for every user as of the latest date both income and rent data are available.

**There are two situation I consider:**

**-1- If rent amount only include the “rent” charge\_type:**

Select income.user\_id as id, rent.amount/income.amount as rent\_to\_income from

(Select user\_id, date, amount,row\_number() OVER (PARTITION BY user\_id ORDER BY date DESC) as rn

From income\_data) income

Join

(Select user\_id, date, amount,row\_number() OVER (PARTITION BY user\_id ORDER BY date DESC) as rn

From rent\_data

where charge\_type = 'Rent') rent

On income.user\_id = rent.user\_id

where income.rn = 1 and rent.rn=1

**-2- if the rent amount include all charge\_types:**

Select income.user\_id as id, rent.rent\_amount/income.amount as rent\_to\_income from

(select USER\_ID,amount from

(select user\_id, date, amount,row\_number() OVER (PARTITION BY user\_id ORDER BY date DESC) as rn

From income\_data) as i

where rn = 1) as income

Join

(select user\_id, sum(amount) as rent\_amount

from

(select \*

from

(select user\_id, charge\_type, date, amount,

row\_number() OVER (PARTITION BY user\_id, charge\_type ORDER BY date DESC) as rn

from rent\_data) as i

where rn = 1 ) as i2

group by user\_id ) as rent

On income.user\_id = rent.user\_id

1. What analysis would you run (without actually coding it up!) to understand the impacts of income against ability to pay rent? Feel free to provide thorough explanation and your thought process.

It’s obvious that income will influence the ability to pay rent, but will it be a simple positive correlation?

* When we consider the impact of income, we need to think of not only the amount of income , but also the **stability of income**: some people have fixed payroll dates, but other people have flexible payroll dates. A guess is people who have fixed payroll dates tend to have better ability to pay rent
* Similarly, when we compare the **payroll date of income and pay rent due day**, we might find that the closer the two dates are, the less likely that people will not pay rent. So we should also consider this factor of income.
* Most important thing when we measure income: use **average income per month** to avoid the influence of payroll frequency difference and fluctuate income.

Next thing to consider is how to **measure the ability to pay rent**, I think of two ways:

* Total paid / total rent amount (the ratio of total pair rent and total rent due): the higher, the better
* Past due ratios (past due times/ total leasing months): the lower, the better

We have analyzed our income factors, and the way to measure the ability to pay. In the next steps, we can write codes on impacts of income against ability to pay rent.

1. Situation: You meet with Karen, our VP of Product, David, our CEO, and Joey, our Head of Data & Analytics, to review your analysis. It's clear that income and the ability to pay rent are correlated, and we want to provide more insight to our landlord partners than a univariate correlation.

You're sent on a discovery project to find what else might be interesting. Write 1-2 paragraphs explaining the question(s) you would ask yourself and/or stakeholders to dig deeper, the approach you would take to discovery, and what data you would test first (assuming it's all at your disposal)?

I need to figure out what is the **most important consideration** about the rent for the stakeholder. So I would ask for example:

which way is better for them:

people pay rent quicker( pay on time regardless of amount) or to pay more (pay total amount eventually but may not on time) to get rent amount quicker or getting more rent amount?

The reason that I ask about this is I need to know what the landlord partners care the most.Our goal is to reduce the risk that people cannot pay their rents. But if they do, which way is relatively unacceptable. That will definitely influence our analysis and model.

Then, there are **some features** **except income** that influence their ability to pay rent we need to consider.

* Type of Household
* Family members (kids, The number of people who can make money)
* Loans
* Class of workers
* Income source
* Marital status
* Number of Owned vehicles

## **Problem #2**

Dataset:

[hybrid\_vehicle\_price\_data.csv](https://s3-us-west-2.amazonaws.com/secure.notion-static.com/7dabe40e-4ed4-4aca-9b15-b7bdc0b20ef6/hybrid_vehicle_price_data.csv)

Please download this dataset for the following exercise. It contains a sample of hybrid vehicles and their characteristics related to price, classification, and performance. See the data dictionary below.

[Data Dictionary](https://www.notion.so/93884bd732f1414ab4e6d8057929aacb)

## **Exercise**

Using R, Python, or another statistical package, perform a regression on this data to model MSRP (price) as a function of other inputs from the data, and then answer these questions. Please share your code and full analysis.

Though not a requirement, consider using R Markdown or Jupyter Notebook, to annotate your work. Another option would be to thoroughly comment in-file.

1. What does your model tell you about hybrid vehicle pricing? What insights can you draw?

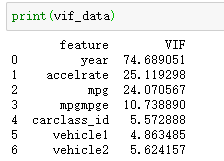
Combining the result from regression models that I choose, I found that:

* Release Year, vehicle model, mpgmpge and accelerate have most important contributions to msrp.
* The newer the release year of the car is, the higher msrp will be.
* The smaller mpgmpge the car has, the higher msrp will be.
* The larger accelerate the car has, the higher msrp will be.

1. Did you use a standard linear regression model or a different type, and if so, why?

I won’t choose a standard linear regression model, since according to my exploratory data analysis : There is high multicollinearity between many features

The reason is the VIF I test as below: It proves high multicollinearity in the features (each higher than 5)



However I tried to use a linear regression model, the performance is not good.

Then I choose to use Tree-based Regression methods like Random Forest, XGBoost, GBM, and LGBM. Benefits of Tree-based models here are :

* They can handle more complex relationships between the features and the output variables.
* They also work well when there is a mix of categorical and numerical features and when there is a big difference in the scale of features.

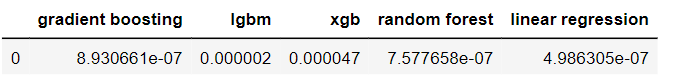
My final decision on the model is the Gradient boosting Model, which is typically optimal for small datasets like in this particular problem. GBM runs sequentially and can choose classifiers from all the features in a step.

Also, K-fold splitting is implemented in GBM. This way, it will reduce variance of the model and avoid overfitting.

The random forest and XGBoost algorithms are parallelizations of the normal decision tree structure. GBM also combines multiple trees but starts at the beginning, and each classifier in a step is predicting the residual of the last step. As opposed to random forest and XGBoost, GBM runs sequentially and can choose classifiers from all the features in a step.

The metrics I used here to determine the accuracy of the model on the validation set was mean squared error (MSE), which measures the average squared difference between the values predicted by our model and the actual values. Comparing different models, the Gradient Boosting Machine has the smallest MSE that means **GBM is the most fitting model** for the dataset.

However, there is no significant difference in MSE between the tree-based models. Below is the comparison between the models that I use.

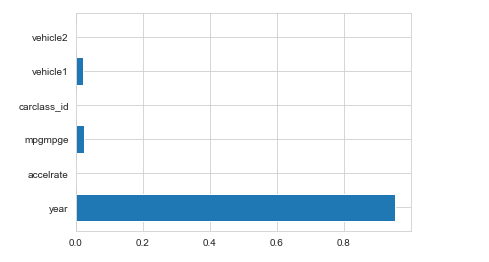


1. Which variables did you choose to include and exclude from your model and why? Did you define any of your own explanatory variables from the data?

After going through the dataset and data dictionary. First, I drop ‘carid’ because this column is only a unique identifier for the vehicle record, and won't influence the price. Then, I drop the ‘carclass’ because the ‘carclass\_id’ and ‘carclass’ are basically the same. So I only keep the ‘carclass\_id’. I also drop the ‘mpg’, because there is correlation between ‘mpg’ and ‘mpgmpge’. I test the VIF and the ‘mpg’ is higher than ‘mpgmpge’ so I drop the ‘mpg’.

Yes, I define my own explanatory variables based on vehicle names. I split the vehicle name into two parts: first part contains model name, second part contains additional information on the vehicle (it could be different generations or different size).

1. Which variables have significant explanatory power on hybrid vehicle pricing? Which do not? Is this surprising?



I export feature importances from each model The ‘feature importances’ were extracted from the trained model, and were used to identify the most important features in predicting user conversion.

These results indicated year, vehicle 1(first part contains model name), mpgmpge were overwhelmingly the most significant predictors of price. Other significant predictors included class\_id, accelerate, and vehicle 2, although these effects were to a much lesser extent than year and vehicle 1.

1. How confident are you in your model's fit? Why?

I’m not very confident in my model fit since the provided dataset is small and includes limited features (features also have high multicollinearity, VIF over than 5).

I choose the GBM because the MSE is smallest and it suits with small datasets. However, it may result in high variance and overfitting in training data.

With bigger datasets, I’m confident that my model will fit better.