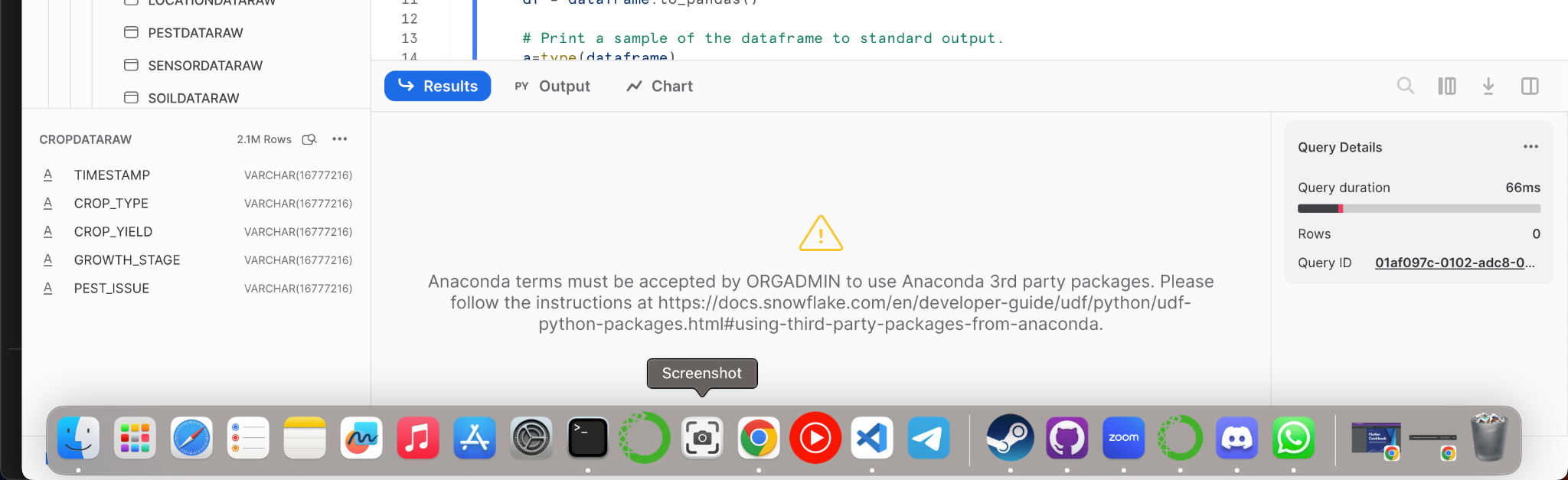
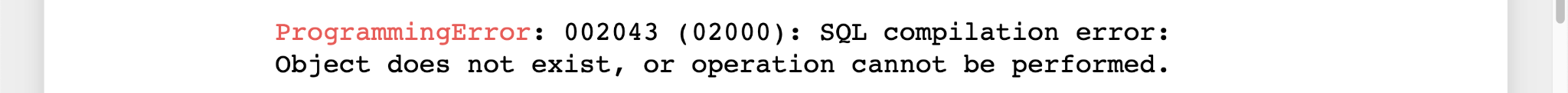
Our data engineering process was as follows:

* + There was a failed Attempt to write python codes on snowflake.
  + Our first attempt was to extract the code using python and transform before loading into the vanguard database.
  + We hit a permission error.
  + 
  + Then we tried to connect to snowflake from jupyter using snowflake-connector-python but hit a database issue.
  + 
  + So we took the data out of snowflake manually and took it through data transformation using [this notebook](https://drive.google.com/file/d/1yfxla1AkAJ4oUNwLMUVYeK9f3PhIaa5X/view?usp=sharing)
  + Here is how we treated [missing data](https://docs.google.com/document/d/1I3Botmth3sDkk0jLGySTsnwmkcogRhpoYDbzXojHd6s/edit?usp=sharing)
  + We also checked the quality of the categorical data AKA Texts
  + After which we uploaded back into new tables called ‘cropclean’ for example.
  + The tables were created using the “create clean tables” worksheet.
  + Then we loaded the clean data as csv.
  + Then used the “create dimension tables” and “create fact table” worksheet to create dimension and facts tables
  + Dimension tables were loaded in immediately
  + Fact table was created in our notebook and loaded directly
  + Since our pipeline was heavy on loading we didn't optimize queries as it would slow down load time.

[An ideal pipeline without any unnecessary query or command](https://drive.google.com/file/d/1xJx2z8iJrssSpa0XnqwoCK4BEGQR-EuP/view?usp=sharing)

[Our Data model documentation and Data Dictionary](https://docs.google.com/document/u/0/d/1G-ZldICPVJr2rMfvyTARW9VqkiesAspeDMit1FVZBts/edit)