# SQL Project Presentation

5 min presentation of approach and findings

1) Data exploration / cleaning
2) Findings from Starting with Questions and Starting with Data
3) Thoughts on future exploration with more time

# Exploration / cleaning

*column_data_inspection.sql, column_cleaning_commands.sql*

- **Null check all columns.** Identify appropriate data types. Dropping entirely empty columns.
  - E.g. Date::DATE – allow MIN(Date) and MAX(Date) to see timeline

- All lowercase column entry names to ease querying, no quotation marks in queries needed.
  *ALTER TABLE products RENAME COLUMN "SKU" TO sku;*

- Transaction related columns e.g. revenue, totaltransactionrevenue divided by 1,000,000
  - E.g. Google Rucksack for 69.99 USD or 69990000 “USD”?

- Finding a **strong indicator** that a transaction occurred. Transactions, ‘Order Confirmation’ Webpage
  - All_sessions.Transactions = 1

- Trimming and concatenating of product names, including Product variants to remove mostly empty data columns. Helps address products that seemingly have multiple SKUs.
  - “ Google Sunglasses “ → ”Google Sunglasses Red”; “ Men’s Hero  Tee” → “Men’s Hero Tee MD”

# Exploration / cleaning

*column_data_inspection.sql, column_cleaning_commands.sql*

- Some erroneous Country entries e.g. Los Angeles, Australia or Mountain View, Japan. City column was taken to be correct.

- Removing redundancy in productcategory e.g. **Home/**Accessories/Fun and **Home/Shop by Brand/**YouTube

- Several null-ish product categories e.g. (not set) which were changed for popular products with large ordered quantities and known category e.g. Nest products and Google Kick Ball

- **QA:** Having created table copies and made updates to the copies I ran NULL checks, expecting a change in record count.

# Exploration / cleaning

*column_data_inspection.sql, column_cleaning_commands.sql*

- Assumptions of data type e.g. varchar for country, integer for time, date, numeric (15,2) for revenues and prices (.99 cents)

- All lowercase column entry names to ease querying, no quotation marks in queries needed.
  *ALTER TABLE products RENAME COLUMN "SKU" TO sku;*

- Utility of data types - fullvisitorid remaining varchar, date type used for date
  *ALTER TABLE analytics ALTER COLUMN date TYPE date USING date::date;*

- Data type assumptions made where some columns were entirely null, not always correct
  - E.g. all_sessions entity, ecommerceaction_type and action_step were 0 and 1, assumed the same of action_option

- Manual reading of ambiguous column labels e.g. “ratio” in sales_report
  - Solved by viewing other integer data types in entity, ratio is total_ordered / stocklevel

# Findings

*starting_with_questions.sql, starting_with_data.sql*

- US Western cities contributed the most transaction revenue from the confirmed transactions available (Grouping cities by region)

- Products like Nest were consistently popular across regions and are among the highest ticket items. Selling particularly well in winter months, they are also well stocked up on.

- Apparel consistently popular particularly in summer months

- Channels: Transactions were mostly via Organic Search, Direct and Referral. Nest products (good sellers) purchased via Referral and Direct, not Organic Search. Affiliate channel had many page views but no conversions, and very few Social page views from analytics

# Challenges / Next Steps

- Filling more revenue fields – 80,000 records exist with NULL revenue but NOT NULL units_sold and unit_price. Only 15,000 completed revenue fields are in the raw analytics file out of 4.3 million records.

- Why does the data show 156,733 ordered quantities of all products but only $14,281 in total transaction revenue? Where is the rest of the recorded transaction data?

- **Route Forward:** Feature manipulation (e.g. units_sold * unit_price) allowed for more data to be revealed but a lot of potential duplicates in analytics.

- **More Time:** Dig into analytics data to identify where more recorded transactions exist. This will help populate the user_sessions total transaction revenue and give more intelligence to the demographics of users purchasing goods on the ecommerce website.