# Lab Exercise 1: Introduction to Data Environment

## Section 1.1 – Accessing Lab Environment

Using putty to access VM

## Section 1.2 – Database Environment – Retail Data

### Step 4

Part 2 – Number of Tables in the training1 database / ddemo schema – **29**  
Part 6 – customers\_dim Table Information

|  |  |  |
| --- | --- | --- |
| **Column Descriptions** | **Type** | **Distribution Key Column(s)** |
| customer\_id  first\_name  last\_name  gender | Integer character varying(100) character varying(200) character(1) | key |

### Step 5

Part 2 – Number of Female Customers – **499041**  
Part 3 – Number of Male Customers – **500959**  
Part 4 – Total Number of Customers – **1000000**

### Step 6

Part 4 – Average expenditure by gender Male: **33.8456101189216746** | Female: **33.7982853835904994**

### Step 7

Top five product categories ordered by men and women

|  |  |  |
| --- | --- | --- |
|  | **Men** | **Women** |
| **1** | Kitchen (28M) | Kitchen (27.6M) |
| **2** | Home Improvement (27.2M) | Home Improvement (27M) |
| **3** | DVD (17.1M) | DVD (17.3M) |
| **4** | Sports (13.6M) | Sports (13.6M) |
| **5** | Automotive Parts and Accessories (10.9M) | Automotive Parts and Accessories (10.8M) |

## Section 1.3 Database Environment – Census Data

### Step 1

Number of Tables in training2 database / public schema – **11**

### Step 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Schema | Name | Type | Owner | Storage |
| public | bayes\_test | table | gpadmin | heap |
| public | fips | table | gpadmin | heap |
| public | foo | table | gpadmin | heap |
| public | housing | table | gpadmin | heap |
| public | housing\_nodupes | table | gpadmin | heap |
| public | income\_state | table | gpadmin | heap |
| public | logr\_coef | table | gpadmin | heap |
| public | nbtrain | table | gpadmin | heap |
| public | persons | table | gpadmin | heap |
| public | zeta | table | gpadmin | heap |
| public | zeta1 | table | gpadmin | heap |

### Step 3

The database has housing information, bio\demographic information, education, and income on people. There seem to be some analytical support tables as well such as “no duplicates (nodupes), bayes\_test, foo, etc.

### Step 4

|  |  |
| --- | --- |
| Name | Rows |
| bayes\_test | 14 |
| fips | 55 |
| foo | 52 |
| housing | 12515394 |
| housing\_nodupes | 6257697 |
| income\_state | 52 |
| logr\_coef | 1 |
| nbtrain | 10010 |
| persons | 28542588 |
| zeta | 64076 |
| zeta1 | 64076 |

### Step 5

In housing, there were a total of **12515394** records with **6257697** duplicates and **6257697** unique.

### Step 6

Deleted existing tables, and ran query to check for duplicates. Total\_dupes = 0.

### Step 7

Part 2 – Correlation between hinc and rooms in state 25 = **0.374485423827578**  
Part 4 – R-square of regression line of hinc and rooms in state 25 = **0.140239332659321**

### Step 8

Created output file from housing\_nodupes and removed summary line.

### Step 9

Created output file from persons and removed summary line.