Navarurh Kumar 1

Course Homepage Individual Final Report **SUBMIT TURNITIN ASSIGNMENT**

Submit Turnitin Assignment

Congratulations - your submission is complete! This is your digital recei
copy of this receipt from within the Document Viewer.

« Page 1 »

Author:
Navarurh Kumar

Assignment title:
Individual Final Report

Submission title:
Final Term Project Report

File name:
NXK180010_BUAN6320_FinalTermProject.pdf

File size:
517.12K

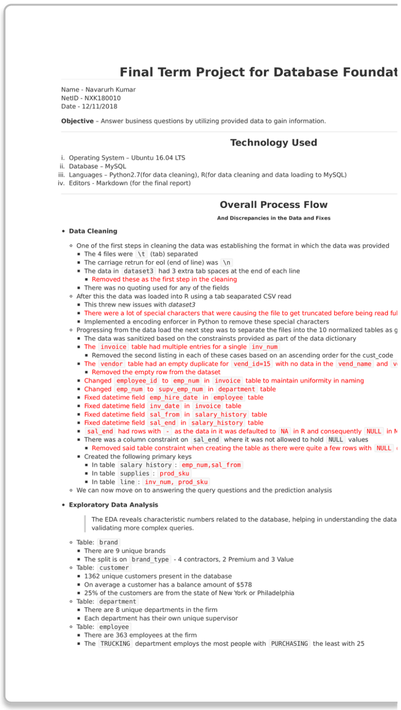
Page count:
17

Word count:
5063

Character count:
24613

Submission date:
12-Dec-2018 08:25PM (UTC-0600)

Submission ID:
1056138766



Final Term Project for Database Foundat

Name - Navarurh Kumar
NetID - NXK180010
Date - 12/11/2018

Objective – Answer business questions by utilizing provided data to gain information.

Technology Used

i. Operating System – Ubuntu 16.04 LTS
ii. Database – MySQL
iii. Languages – Python2.7(for data cleaning), R(for data cleaning and data loading to MySQL)
iv. Editors - Markdown (for the final report)

Overall Process Flow

And Discrepancies in the Data and Fixes

- **Data Cleaning**
  - One of the first steps in cleaning the data was establishing the format in which the data was provided
    - The 4 files were `\t` (tab) separated
    - The carriage retrun for eol (end of line) was `\n`
    - The data in `dataset3` had 3 extra tab spaces at the end of each line
      - Removed these as the first step in the cleaning
    - There was no quoting used for any of the fields
  - After this the data was loaded into R using a tab seaparated CSV read
    - This threw new issues with *dataset3*
    - There were a lot of special characters that were causing the file to get truncated before being read ful
    - Implemented a encoding enforcer in Python to remove these special characters
  - Progressing from the data load the next step was to separate the files into the 10 normalized tables as g
    - The data was sanitized based on the constrainsts provided as part of the data dictionary
    - The `invoice` table had multiple entries for a single `inv_num`
      - Removed the second listing in each of these cases based on an ascending order for the cust_code
    - The `vendor` table had an empty duplicate for `vend_id=15` with no data in the `vend_name` and `v
      - Removed the empty row from the dataset
    - Changed `employee_id` to `emp_num` in `invoice` table to maintain uniformity in naming
    - Changed `emp_num` to `supv_emp_num` in `department` table
    - Fixed datetime field `emp_hire_date` in `employee` table
    - Fixed datetime field `inv_date` in `invoice` table
    - Fixed datetime field `sal_from` in `salary_history` table
    - Fixed datetime field `sal_end` in `salary_history` table
    - `sal_end` had rows with `-` as the data in it was defaulted to `NA` in R and consequently `NULL` in M
    - There was a column constraint on `sal_end` where it was not allowed to hold `NULL` values
      - Removed said table constraint when creating the table as there were quite a few rows with `NULL`
    - Created the following primary keys
      - In table `salary history` : `emp_num,sal_from`
      - In table `supplies` : `prod_sku`
      - In table `line` : `inv_num, prod_sku`
  - We can now move on to answering the query questions and the prediction analysis
- **Exploratory Data Analysis**

  > The EDA reveals characteristic numbers related to the database, helping in understanding the data
  > validating more complex queries.

  - Table: `brand`
    - There are 9 unique brands
    - The split is on `brand_type` - 4 contractors, 2 Premium and 3 Value
  - Table: `customer`
    - 1362 unique customers present in the database
    - On average a customer has a balance amount of $578
    - 25% of the customers are from the state of New York or Philadelphia
  - Table: `department`
    - There are 8 unique departments in the firm
    - Each department has their own unique supervisor
  - Table: `employee`
    - There are 363 employees at the firm
    - The `TRUCKING` department employs the most people with `PURCHASING` the least with 25

We take your privacy very seriously. We do not share your details for marketing purposes with any externa
be shared with our third party partners ONLY so that we may offer our service.

Return to assignment list