**The British College**

**KATHMANDU**

**Coursework Submission Coversheet**(individual coursework only)

**Faculty of Arts, Environment and Technology LBU Student Id:**

77178367

**For checking by the student:**

Please ensure all information is complete and correct and attach this form securely to the front of your work before posting it in a coursework collection box.

Award name: BSc (Hons) Computing

Module code: 23269 – 201516

Module name: Advanced Database A

Module run: 2016

Coursework title: 2

Due Date: 19 November (2016)

Module leader: (In LBU) Jackie Campbell, Sanela Lazarevski

Module tutor: (In TBC) Dibya Tara Shakya

**TURNITIN** Checked: YES NO ***(please circle)***

Submission date& time: Date: 19 September (2016) Time: 10 PM

**Total Word Count: Total Number of Pages (including this front sheet):**

**In submitting this form with your assignment, you make the following declaration:**  
I declare, that the coursework submitted is my own work and has not (either in whole or part) been submitted towards the award of any other qualification either at LBU or elsewhere. I have fully attributed/referenced all sources of information used during the completion of my assignment, and I am aware that failure to do so constitutes an assessment offence.

Signed: Date:

**You are strongly advised to retain a second copy of your work in case of any query about the assignment.**

3327

30

**For completion by the faculty:**

**This mark is provisional and subject to moderation and approval by the relevant examining board**

**Teacher's Feedback**

**Teacher's Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Contents

[Task 1 A 3](#_Toc467331797)

[ETL (Extract, Transform, Load) Script 3](#_Toc467331798)

[Creating Tables 4](#_Toc467331799)

[Managing Relations 9](#_Toc467331800)

[TEMP\_VEHICLETYPE 9](#_Toc467331801)

[LOAD\_ERRORS 10](#_Toc467331802)

[DATA\_QUALITY\_ERRORS 11](#_Toc467331803)

[DATA\_TRANSFORMATIONS 11](#_Toc467331804)

[TRANSPORTSTAGE\_TEMP 12](#_Toc467331805)

[Altering Table 13](#_Toc467331806)

[PACKAGE\_SPEC 14](#_Toc467331807)

[Executing Procedures 14](#_Toc467331808)

[ERRORS 17](#_Toc467331809)

[ETL Evidence 18](#_Toc467331810)

[Fact Table Evidence 18](#_Toc467331811)

[Error Table Evidence 19](#_Toc467331812)

[Data Transformation Evidence 19](#_Toc467331813)

[Summary 19](#_Toc467331814)

[Task 1B 20](#_Toc467331815)

[Basic SQL Report 20](#_Toc467331816)

[Report 1 20](#_Toc467331817)

[Apex Interactive Reports 21](#_Toc467331818)

[Report 2 21](#_Toc467331819)

[Report 3 23](#_Toc467331820)

[Task 2 A 25](#_Toc467331821)

[Report 1 25](#_Toc467331822)

[Report 2 26](#_Toc467331823)

[Task 2 B 28](#_Toc467331824)

[Introduction 28](#_Toc467331825)

[Use of Data Mining 28](#_Toc467331826)

[Data Mining Techniques 29](#_Toc467331827)

[Association Technique 29](#_Toc467331828)

[Bibliography 29](#_Toc467331829)

# Task 1 A

# ETL (Extract, Transform, Load) Script

To perform Extract, Transform and Load also known as ETL work for Transport Department, data available with city council requires to be gathered into a temporary stage server (stage table) first. City council has provided various record files in CSV format. All records are first transferred into single CSV file then loaded into staging table using APEX tool. Here information from all files transferred into single file is for the ease during upload only. All files can be uploaded one at time as well. This process is understood as extract.

Stage Table

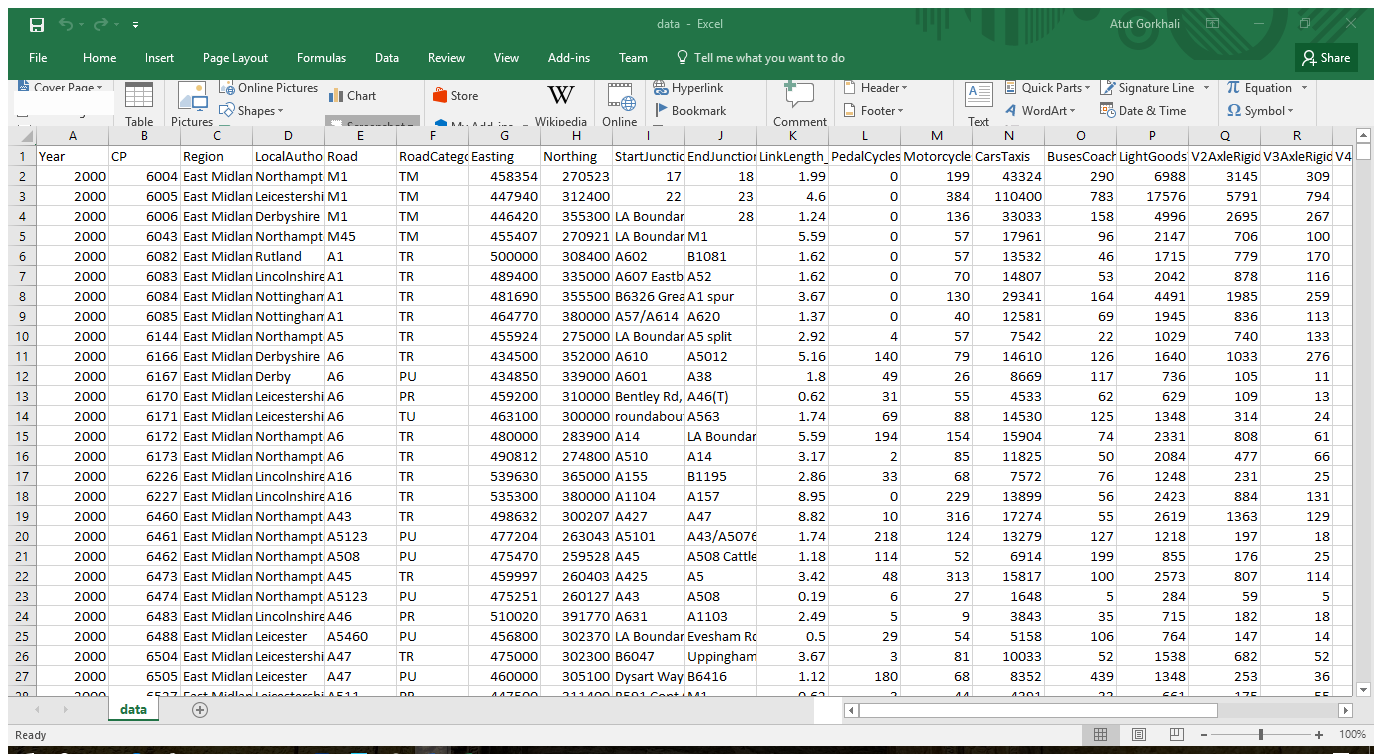


Figure 1Available data with City Council

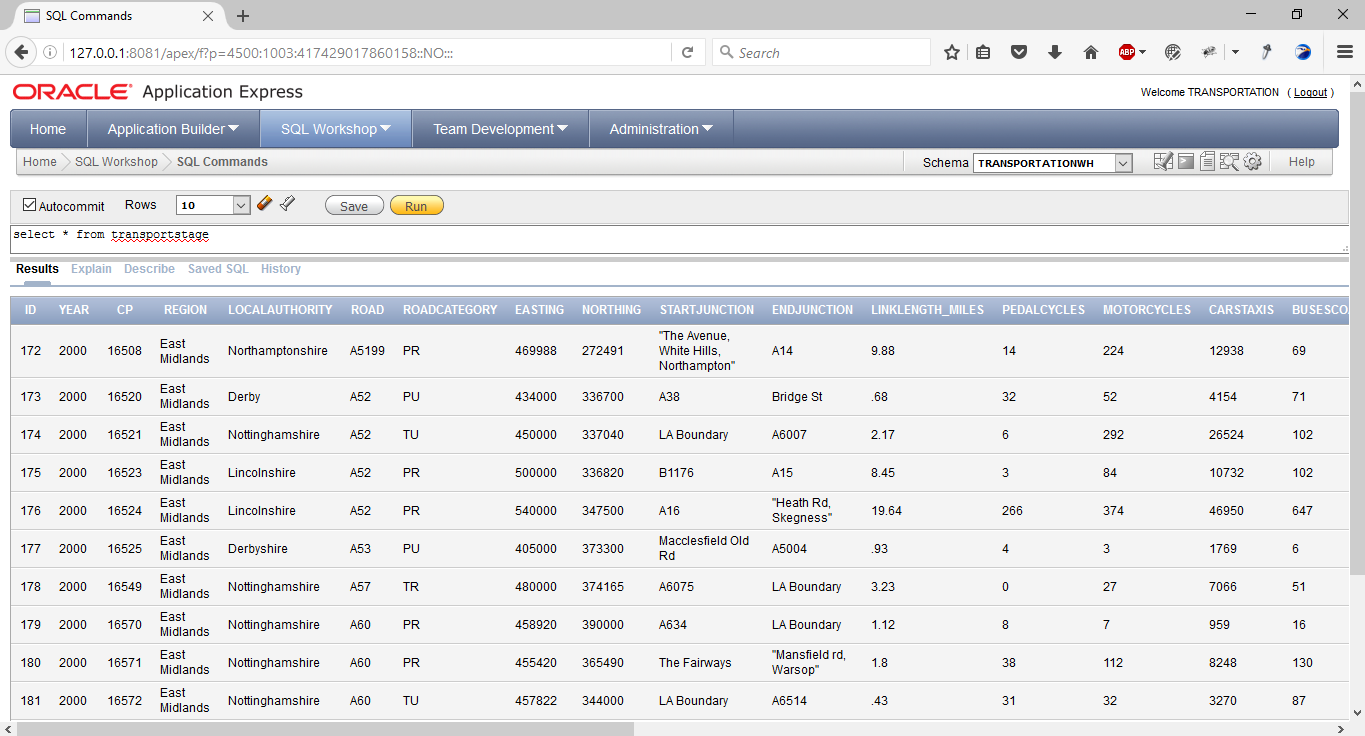
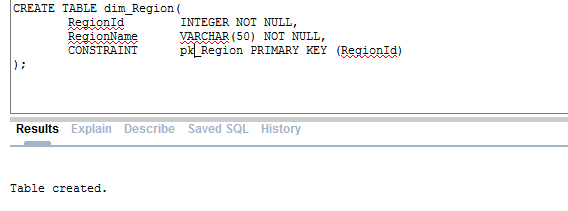


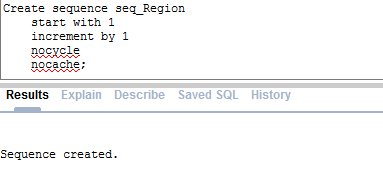
Figure 2Data Loaded into APEX

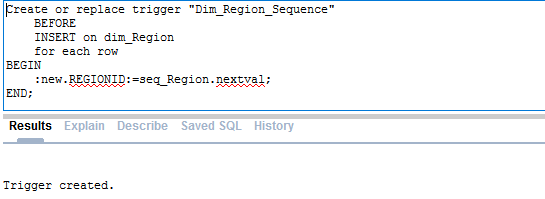
# Creating Tables

To develop warehouse system for city council first various dimension tables and fact table created. Scripts for developing dimension tables and fact tables and evidence of script execution is provided below. Each dimension tables and fact table contains primary key that is designed to be incremented automatically each time new information is inserted. For this, separated sequence and trigger for each table to increment sequence on each insert event. Information from available data will be loaded into these tables.

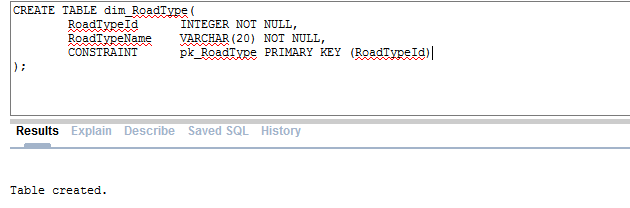
DIM\_REGION

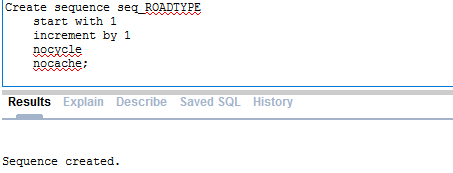


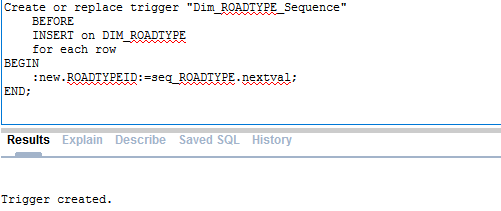




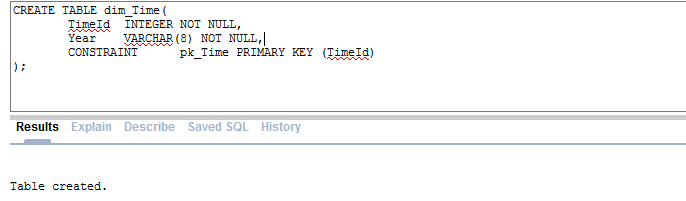
DIM\_ROADTYPE

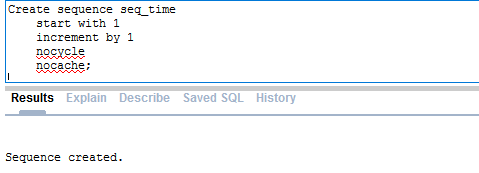


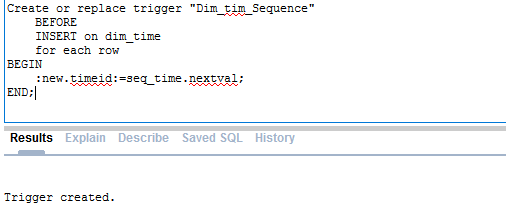




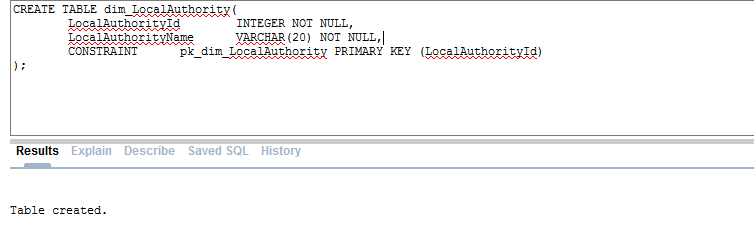
DIM\_TIME

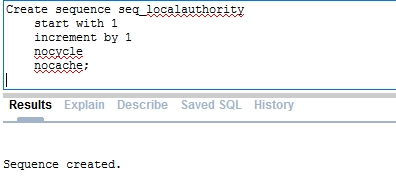


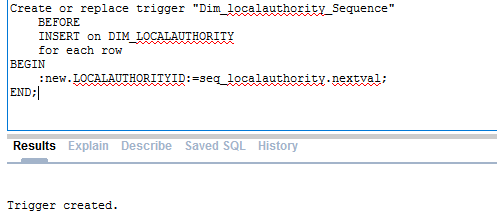




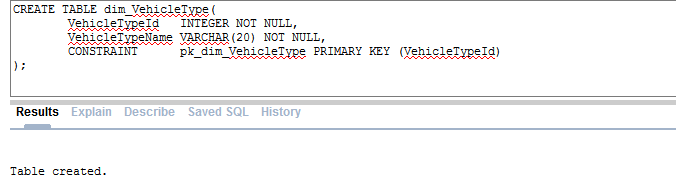
DIM\_LOCALAUTHORITY

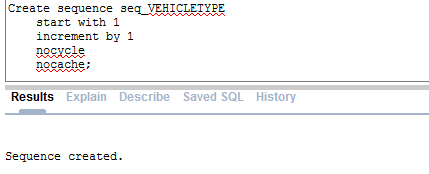


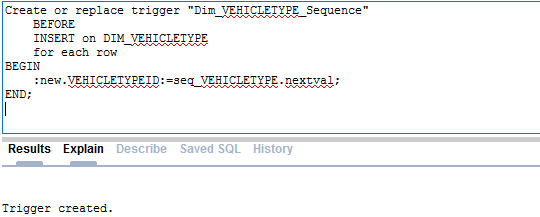




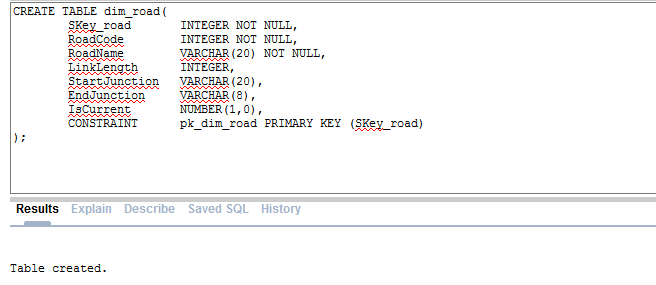
DIM\_VEHICLETYPE



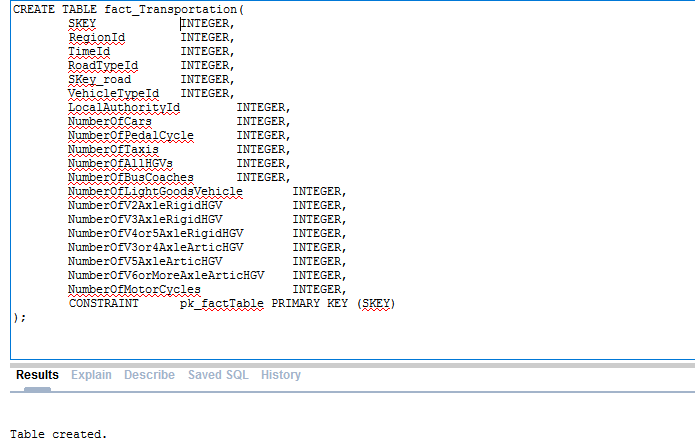


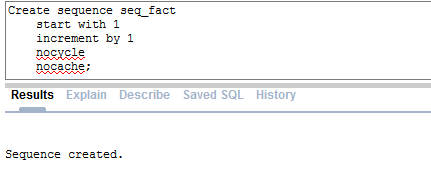


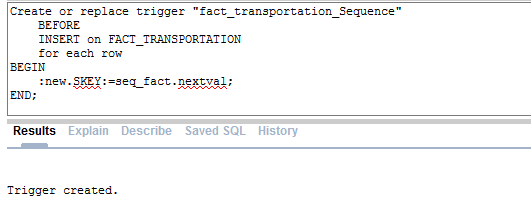
DIM\_ROAD



FACT\_TABLE

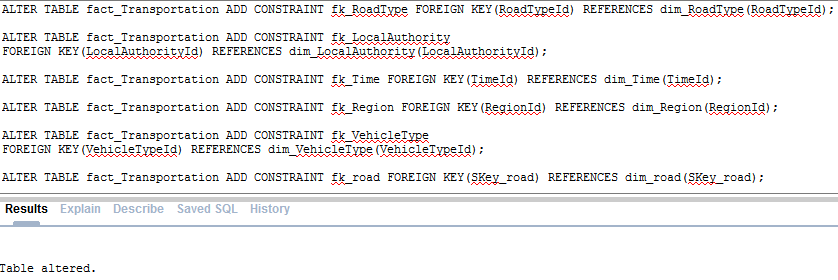






## Managing Relations

Now as each of the required tables for warehouse is created, their relation with other tables is need to be maintained. This is achieved by creating relative using primary and foreign keys between tables. Evidence of relationship maintenance is provided below.



## TEMP\_VEHICLETYPE

Vehicle types in available data is not maintained in a way that can be loaded into dimension as shown in figure 3 below. Hence a table is created named TEM\_VEHICLETYPE on which vehicle types will be inserted manually. Vehicle types in this table would later then cleaned and loaded into vehicle dimension table.

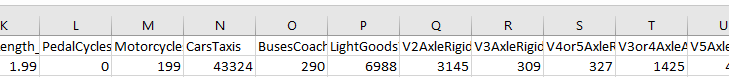
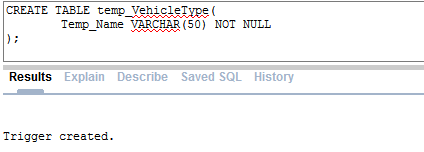


Figure 3 How vehicles are maintained in available information



After creating temp vehicle table useful vehicle types are inserted manually using script show below.

Insert Into TEMP\_VEHICLETYPE(TEMP\_NAME) Values('PEDALCYCLES');

Insert Into TEMP\_VEHICLETYPE(TEMP\_NAME) Values('MOTORCYCLES');

Insert Into TEMP\_VEHICLETYPE(TEMP\_NAME) Values('CARSTAXIS');

Insert Into TEMP\_VEHICLETYPE(TEMP\_NAME) Values('BUSESCOACHES');

Insert Into TEMP\_VEHICLETYPE(TEMP\_NAME) Values('LIGHTGOODSVEHICLES');

Insert Into TEMP\_VEHICLETYPE(TEMP\_NAME) Values('V2AXLERIGIDHGV');

Insert Into TEMP\_VEHICLETYPE(TEMP\_NAME) Values('V3AXLERIGIDHGV');

Insert Into TEMP\_VEHICLETYPE(TEMP\_NAME) Values('V4OR5AXLERIGIDHGV');

Insert Into TEMP\_VEHICLETYPE(TEMP\_NAME) Values('V3OR4AXLEARTICHGV');

Insert Into TEMP\_VEHICLETYPE(TEMP\_NAME) Values('V5AXLEARTICHGV');

Insert Into TEMP\_VEHICLETYPE(TEMP\_NAME) Values('V6ORMOREAXLEARTICHGV');

After successfully inserting all vehicle types, entry can be confirmed using select statement as shown in figure 4 below.

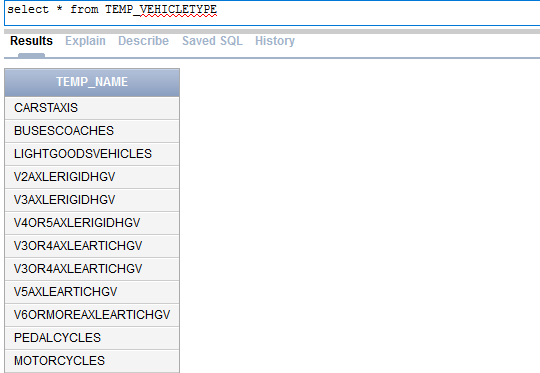
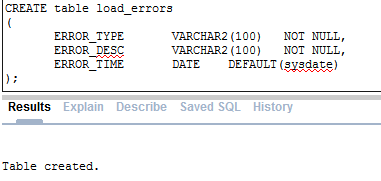


Figure 4 Evidence of manual vehicle type entry

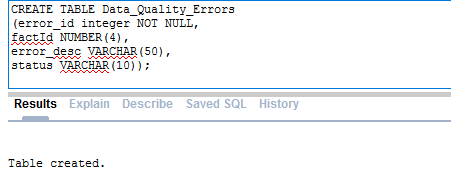
## LOAD\_ERRORS

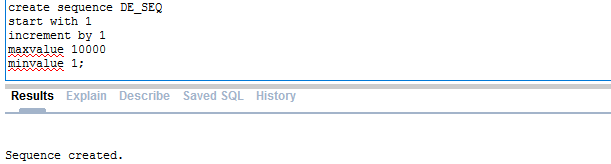
Now the dimension table as well as sequence and triggered required for ETL process has already been created. During the ETL process, not all information can be loaded into warehouse. There is always possibility of some errors occurring. It is essential to track log for each error. The helps to ensure data quality of system and also helps to maintain the system. In this warehouse project, a error table has been created that would automatically store error logs during loading process.



## DATA\_QUALITY\_ERRORS

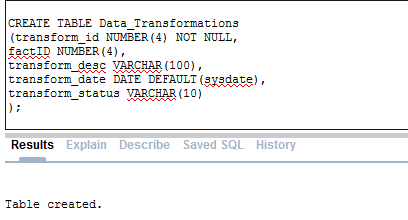
Data quality is essential aspect of ware house. Data quality ensure available information are useful and give necessary information. To maintain the data quality, it is important to clean available information. Hence in warehouse system for city council available data are cleaned first before ware housing. Data with no issue is loaded into a temporary stage table first and data with issues are logged into data quality error table. Tables required for logging quality errors are created below.

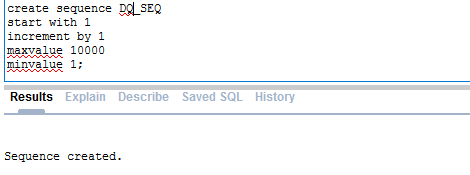


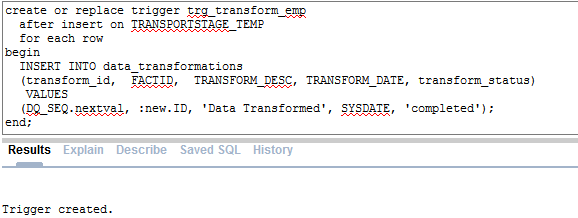


## DATA\_TRANSFORMATIONS

Data cleaning is part of data transformation. During data cleaning, data transformation is logged in a transformation table. Transformation log table and necessary sequence as well as trigger is created below. This process is transformation.

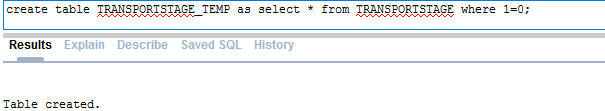






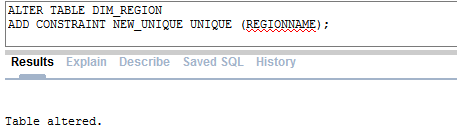
## TRANSPORTSTAGE\_TEMP

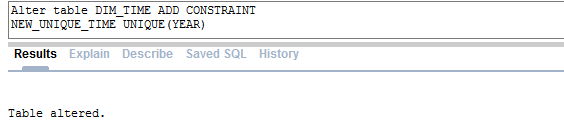
As mentioned above, after cleaning the available data, cleaned data into another staging table. Script for creating temp staging table is shown below.

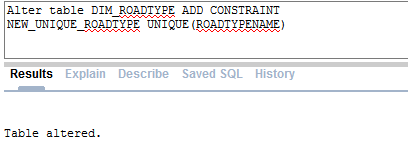


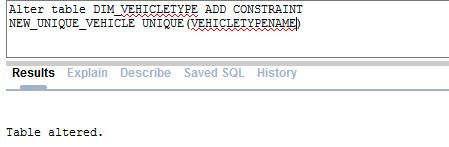
## Altering Table

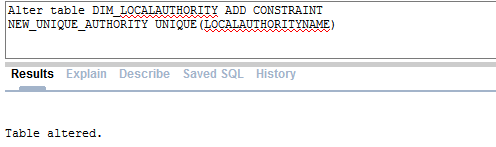
Created dimension table’s attribute are made unique. This means there can be more than one region, road, vehicle type or year with same name. This issue is solved by altering the table and adding unique constraint. Scripts written to alter tables are as follows.

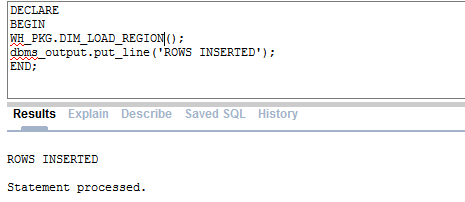






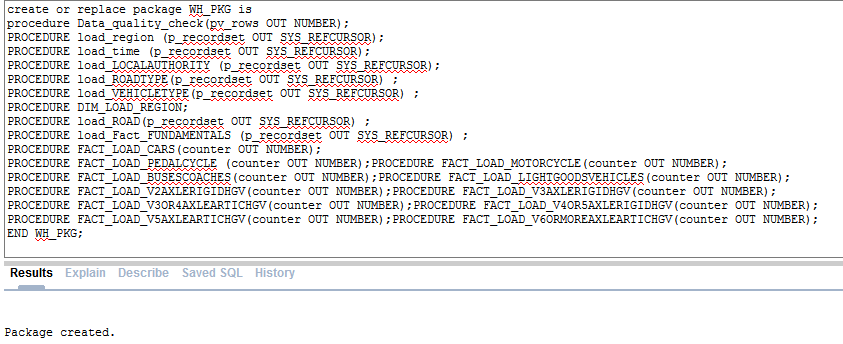






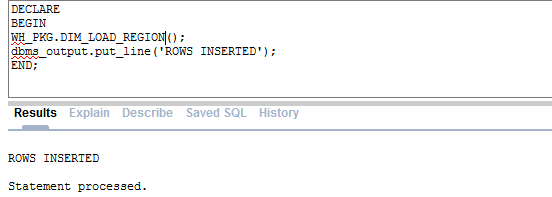
# PACKAGE\_SPEC

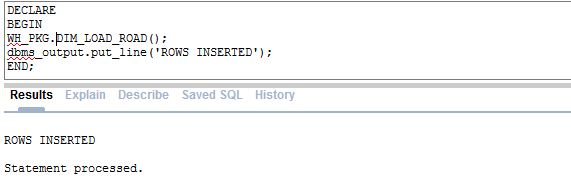
Script of Warehouse project for city council consist of number of procedures. These procedures are maintained in a single package. Package is nothing but grouping of related procedures, functions or variables. A package usually has two parts that are specification and body. Creating package results better performance as when package is called, whole package is loaded into memory hence calling other functions or procedure form same package shows improved performance. Package includes procedures to out cursor of various dimension tables such as vehicle types, road, region and time etc. Specification part of the package which is just interface to the body part is shown below. However, it is not implemented in specification part of package but the body part of package. Both specification and body of package is included in the assignment zip.

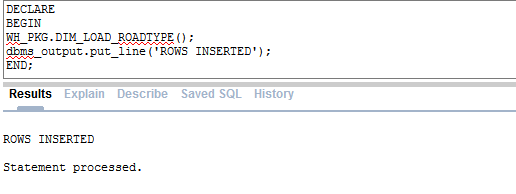


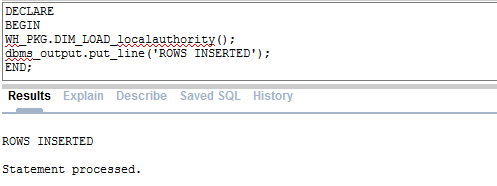
# Executing Procedures

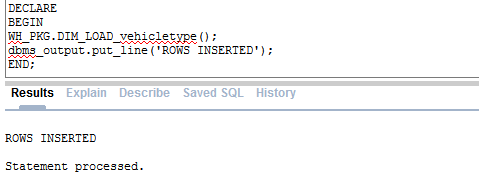
Once each procedure specified in package specification is coded in body part, each procedures are required to be executed. To do this, package name is used along with dot (.) operator to access various procedures with in the package. This process is load process. Execution of all necessary procedure to clean data, load data into dimension and fact tables are shown below.

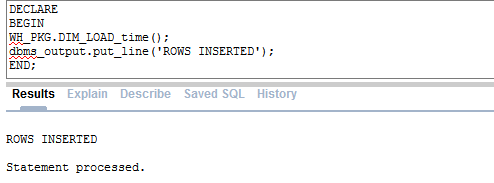


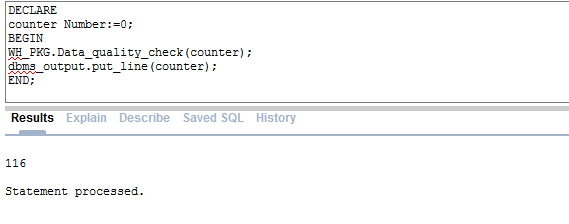


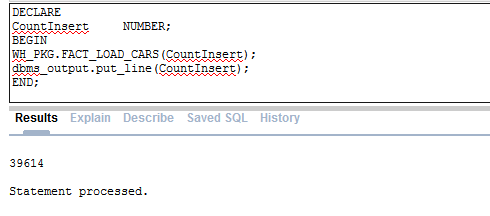






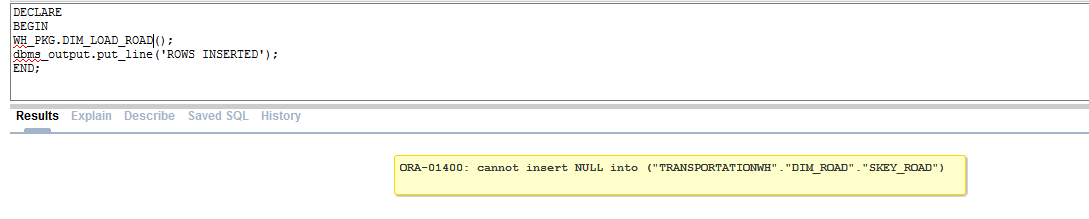




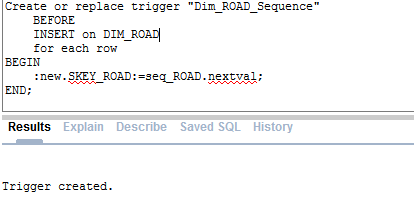


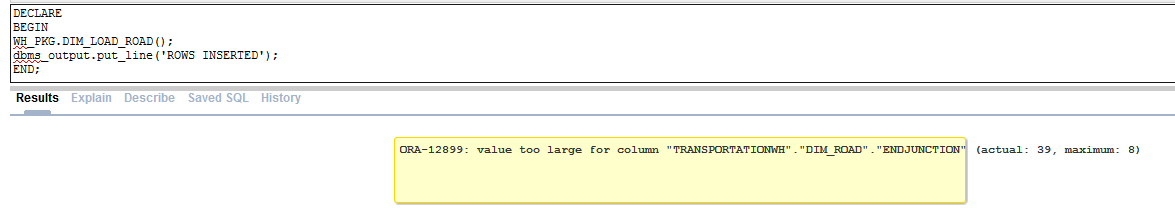
# ERRORS

During the execution of various procedures or creating sequences, triggers or tables, there were few issues arose. Some of the issue where due to coding errors while some of them where due to wrong data types.

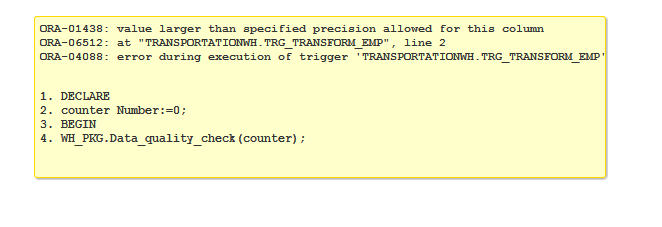


In the issue shown above, while executing procedure to load roads into dimension table it is found that table has no sequence association. Hence sequence and trigger for road dimension table is created and solved the issue.

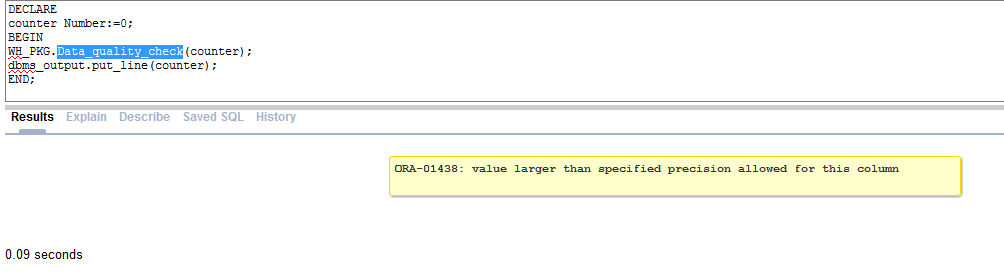




Now after solving the sequence related issue in road dimension table, it was found that END JUNCTION attribute of table can only support up to eight character. Hence to support necessary information, attributes of table where altered and solved the issue.



Another issue was noted when trying to execute data quality check procedure. Error showed size of the variable was smaller than required, hence table was modified and issue was resolved.



# ETL Evidence

Now since all the necessary procedures are executed and error found during the execution process has been resolved. After resolving these errors, resolved procedures were executed as well. Hence, now warehouse must have all loaded information. Data quality table should have logs, transformation table should have transformation logs, fact table must have fact data and error table must have stored error logs. Evidence of successful ETL process is shown below using simple select statements of error, transformation and fact tables.

## Fact Table Evidence

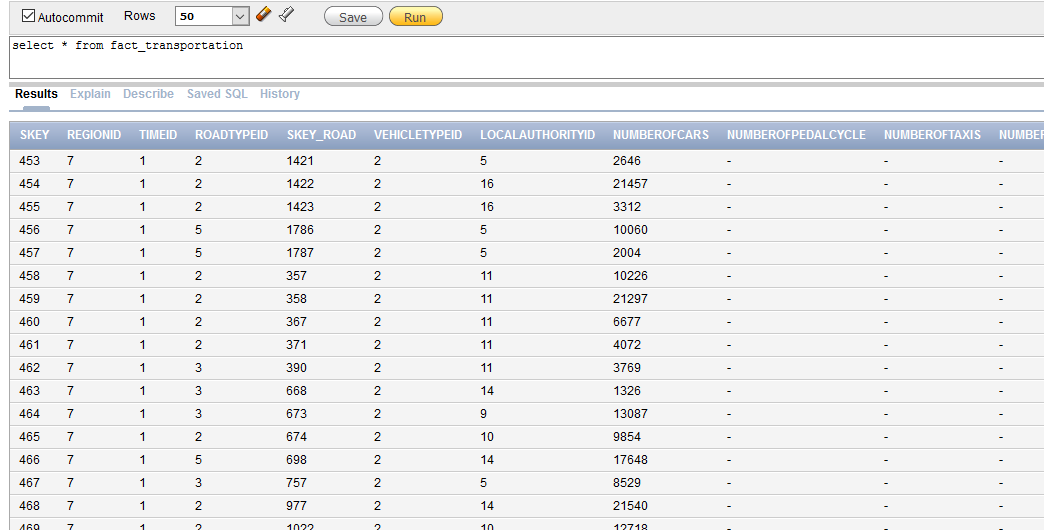
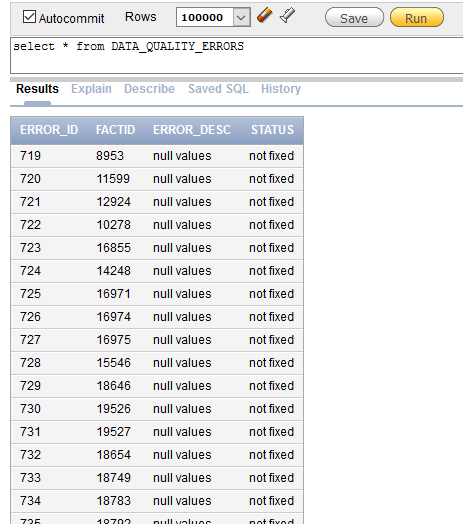
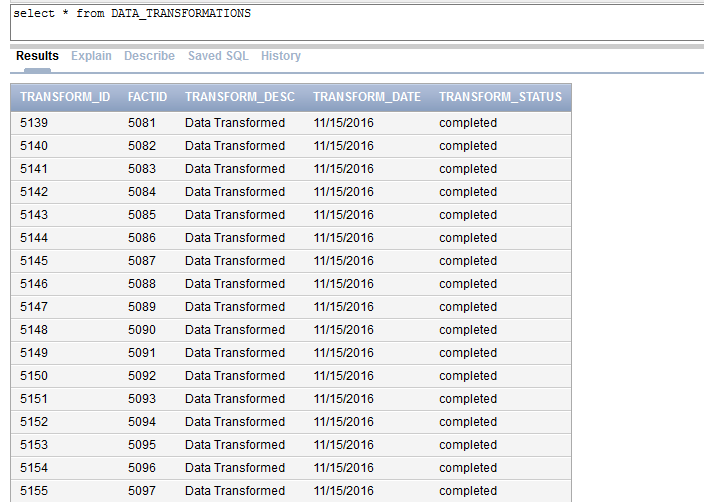


Figure 5 Evidence of data loading in fact table

## Error Table Evidence



## Data Transformation Evidence



# Summary

This document has populated data into warehouse system developed during assignment 1. To populate data, existing data needs to be go through various phases before loaded into actual system. Data requires to be first extracted from source into a system, named stage table in this project. Then it is cleaned by removing data with issues and only taking cleaned and useful information.

Cleaned data is then stored into temp stage table. Then using procedures useful information are stored into dimension and fact tables. Different tests were carried then to ensure successful data load. Unlike OLTP databases, now loaded information will can be utilized to analyze requirement of transport department and help development of city council.

# Task 1B

Generating reports from warehouse is as important as developing them. Transport department of city council has some report requirements that can help to improve the situation of transportation in the region. This document now creates some the reports that are were found required by city council during assignment 1.

## Basic SQL Report

## Report 1

**Generate** report to differentiate on number of bus and number of cars in various local authorities in Year 2010 within Yorkshire Region

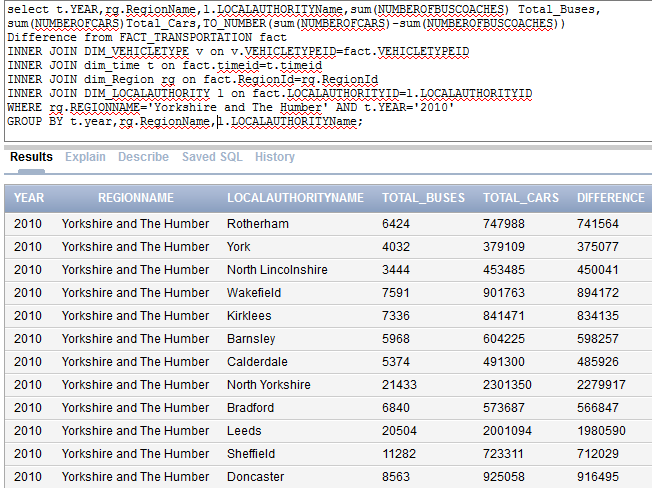


Figure 6Report generation using JOIN SQL statement

Here in figure 6 above, shows the creation of basic SQL report by joining fact and dimension table to produce useful information. Fact table stores keys of various dimension table that is not understandable to non-technical stakeholders. To develop friendly report, fact table is joined with various dimension table and took necessary fields from all tables. In above report number of buses and number of cars are listed within Yorkshire region in year 2010.

Similar reports can be produces on different year this helps to understand vehicle trends. For example, in this report it show cars are more favored than buses. This means city council can develop different policies to promote city buses, this will help to reduce number of car and taxis in city. Eventually such policy would help to make roads less crowed.

## Apex Interactive Reports

Oracle Apex allows to create interactive and user friendly reports. Unlike basic SQL reports, APEX report are more interactive and allows to showcase information in more friendly manner.

## Report 2

**Generate** Report to analyze most car crowded road and car number in different years and regions

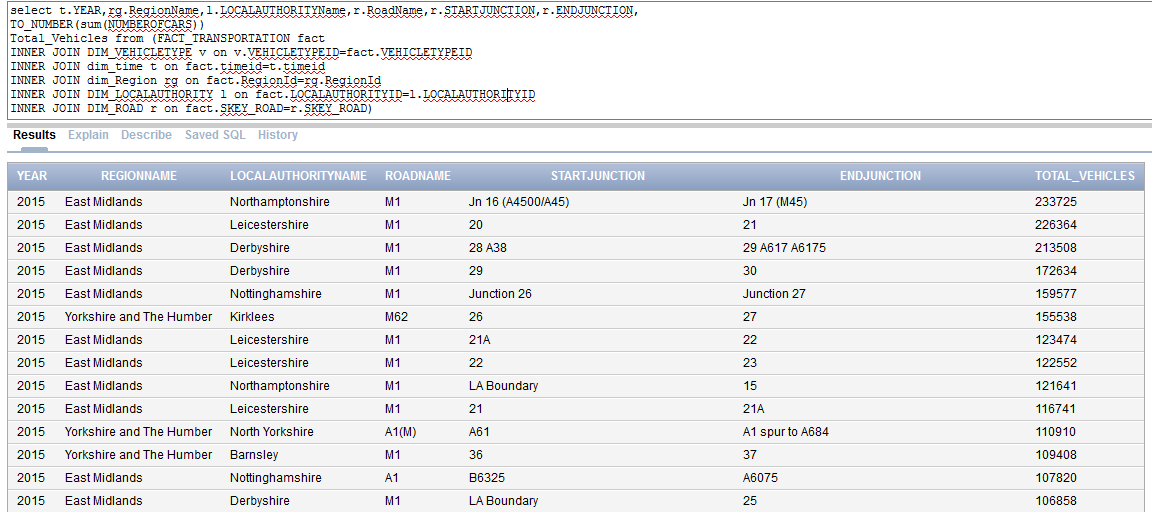


Figure 7Script to generate Apex report

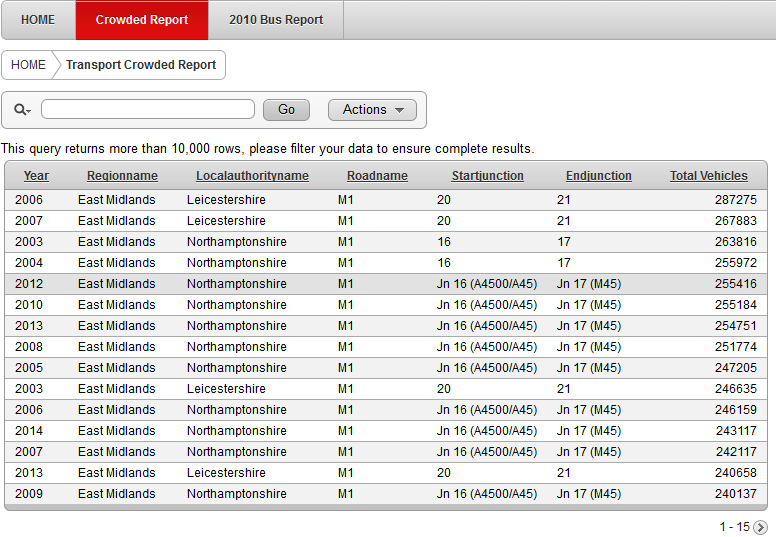


Figure 8 Total Vehicle Report

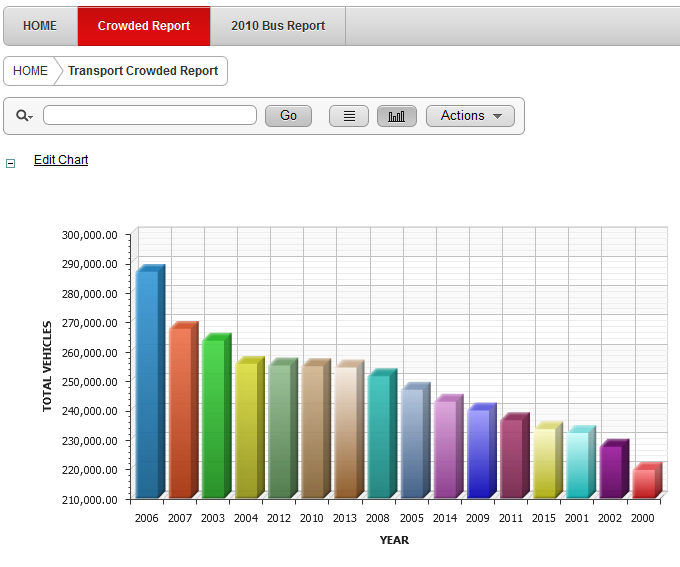


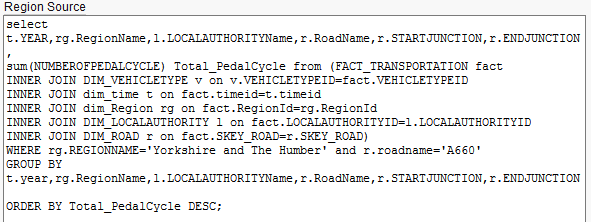
Figure 9 Report on number of car over the year

Using Apex application builder section a report is generated here to satisfy the requirement of city council. One requirement of city council is to report number of car over the years and understand the trend of using car. SQL script written in figure 7 is utilized during creation of apex report.

Figure 9 above demonstrates the created apex report. It shows number of cars and year over the year. Report shows in 2000 car was least used compared in other years while in 2006 most cars were used. Using these reports data expect can analyzed the trend in use of different vehicle types. Unlike table based basic SQL report, APEX interactive report give visual presentation, give ability to search necessary information. User also has ability to recreate report from given information in the page, shown in figure 8. Analyzing trend of vehicle use helps to understand requirement of new roads or development strategies. Report shows year 2006 and 2007 were most crowded year. However recent years such as 2012-2015 has average vehicle crowed compared to most and least crowded year. This shows need of widening the road are not necessary to be in top priority list. However other report should also be analyzed before making development plans.

## Report 3

**Generate** report the number of pedal cycle in ‘A660’ road different year to analyze need cycle friendly roads



SQL join script above is utilized to prepare apex report to analyze number of cycle in road ‘A660’ and numbers throughout the years. First apex table report is created that lists number of pedal cycle in descending order as shown in figure 10. Later same table is used to create horizontal bar report.

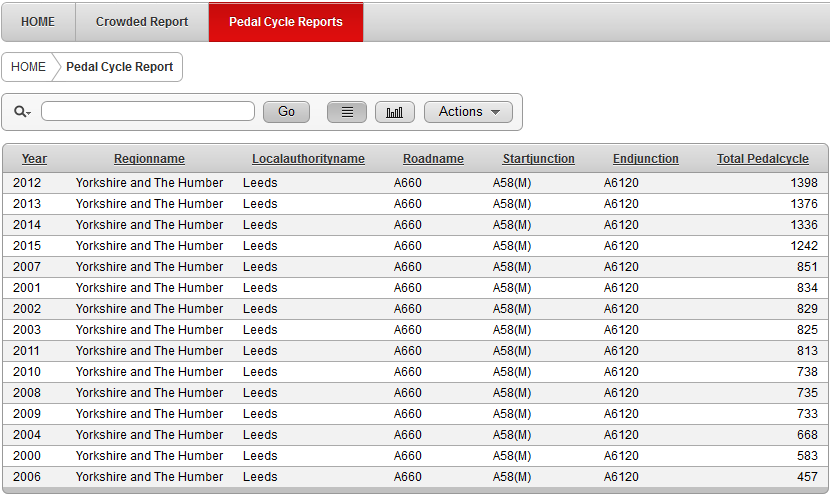


Figure 10 Apex Table report on Pedal Cycle

It shows in year 2012 to 2015 number of pedal cycle has crossed 1000 marks. This can help to analyze the need of cycle friendly road.

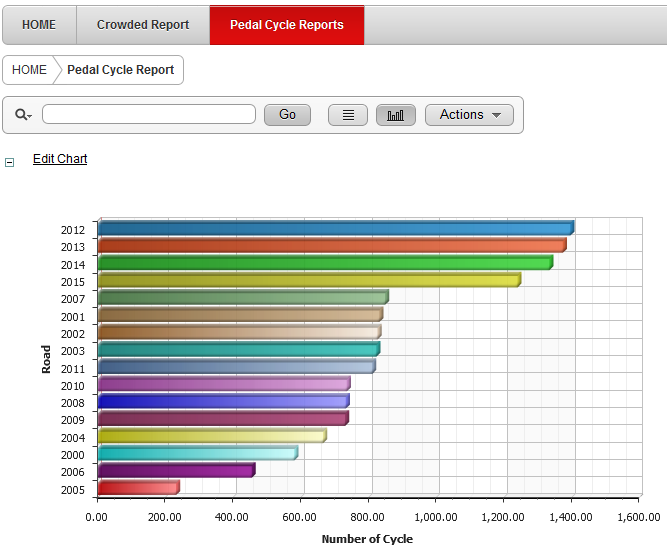


Figure 11 Apex horizontal char on Pedal cycle

Figure 11 shows apex horizontal pedal cycle chart. This friendly chat shows in year 2012 most number of pedal cycles where used. Most recently in 2015 over 1200 pedal cycle ran in road ‘A660’. This report also helps to plan strategies to promote environment friendly transportation.

# Task 2 A

Pivot table is tool that enables analyst to get organized and summarized report from available data in database or spreadsheet. It does not modify the data itself but only manipulate the presentation of data. City council has various data sheet available on transportation. To get better report, first all spreadsheets are combined into a single spreadsheet. Then using Microsoft excel tool, pivot table is created. This document includes two different reports that are requirement of city council and would help to upgrade the standard of transportation facility within region.

Report 1: **Generate** report on Comparison between number of Cars/Taxis, Buses and Light Goods Vehicle

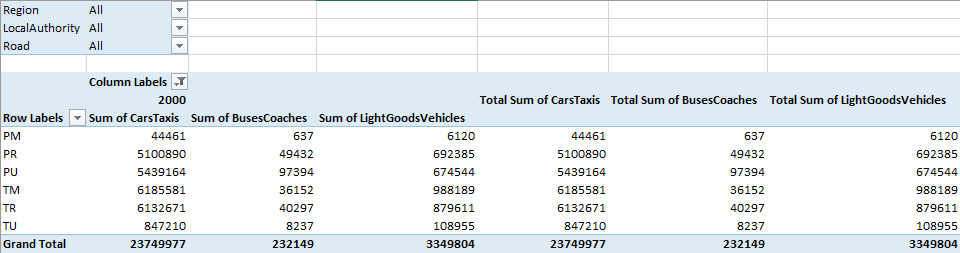


Figure 12 Pivot Table to compare different vehicle type numbers

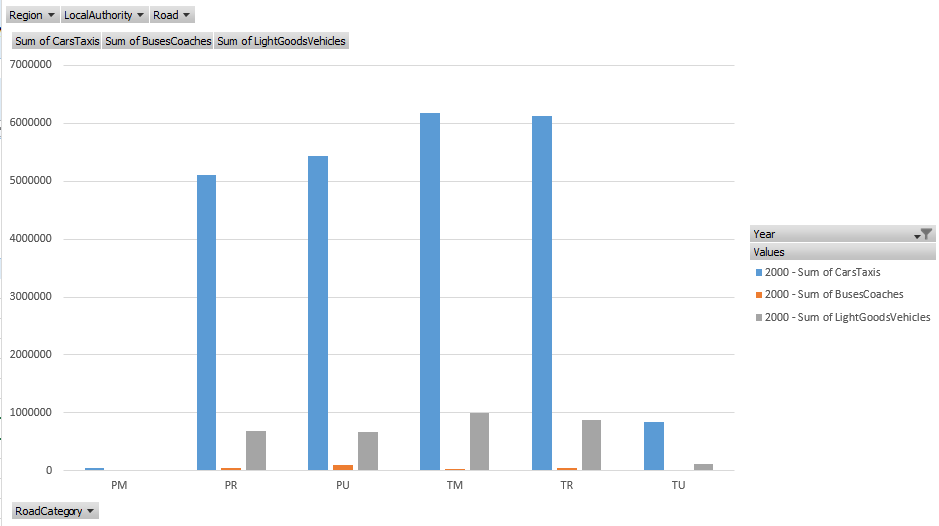


Figure 13Pivot chart comparing different type of vehicles in year 2010

This report would help data experts to analyze the trend of different vehicle types and how they are making impact on different road types. Before making visual presentation of report, first pivot table is created as shown in figure 12. To do so existing data are first transformed into pivot table. A pivot table has four sections: filter, Columns, Rows and Values. Filter section allows to filter the information. Column section specifies column headers while rows specifies rows header. And finally, values section specifies value information. In this report, Filter section contains region, local authority and roads. Column section contains year and header for sum of car and buses. Only 2010 year is selected to filter only 2010 information. Rows section contains road categories and finally values section contains values of sum of car and sum of buses.

Using insert pivot char feature of Microsoft excel then pivot char is created as shown in figure 13. Bar char report shows in 2010 car/taxis are fact more used vehicle than buses or light good vehicles in all region. This report helps decision makers to make plans to balance the mean of transportation.

Report 2: **Generate** report to analyze number of vehicles in all road and regions in year 2015

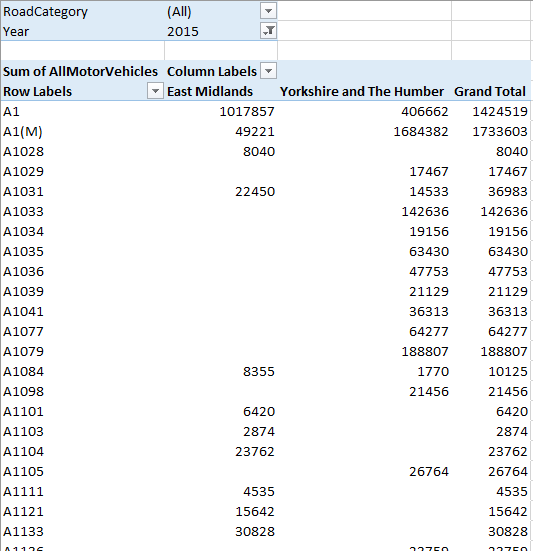


Figure 14 Pivot table to report sum of all vehicles in different road/region

Like in first report, in this report pivot table is created first before preparing a pie char representation of information as shown in figure 14. In pivot table year and road category are placed in filter section to filter the desired data. Here only data of year 2015 calculated. Column section includes region so value of all sum of vehicle in all region is can be compared. Rows section contains road as head hence value can be displayed for each road type. This report also demonstrates which road go through which region. After preparing pivot table, pie chart is created as shown in figure 14.

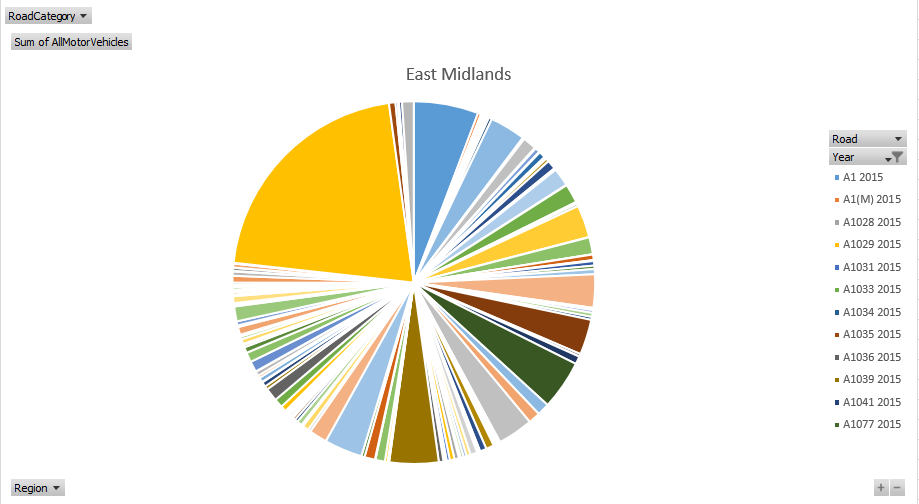


Figure 15 Pie char report on vehicle crowed in different roads in year 2015

Report shows in 2015, A1029 road is most crowded road compared to other roads. A1 road comes in second place. Showcase of information in pie char enables all stakeholders to understand the situation without making much effort. Using such report, strategy planner would able to plan development strategies for upcoming year hence over crowded road does not have to suffer long traffic. Overcrowded road usually requires more maintenance as well. Hence such reports help to develop transportation of region.

# Task 2 B

## Introduction

Data itself is not understandable to all stakeholders and are not useful unless it is transformed into information. This transformation of data to information can be understood as data mining. In simple word data mining is analyzing data from various viewpoint extract useful information. Even though data mining is newer term, it itself is not new though. Organizations has utilized this technology to understand trend or public demand over the years. There are three major aspects of data mining and they are data, information and knowledge.

As discussed above, data itself cannot be used to analyze and predict or understand pattern. Data can be anything such as sales record, payroll record or in this project case number of different vehicle are data. These data are then transformed into information. For example, in this current project data are transformed into information to understand total number of vehicle in different years, most used vehicles and most crowded road etc. This helps to transform Information into knowledge. Here, knowledge means understanding pattern of vehicle or road use so that transportation department can make plans to develop the region.

## Use of Data Mining

Some of the use of data mining technologies are market study, science study, predictions and production management. Leeds City council can use data mining technology to understand trend of vehicle use and pattern of road user etc. Following are some of the areas where data mining can be used for development of transportation.

1. Vehicle profiling

Understand what type of vehicle are preferred

1. Identify road requirements

Understand the requirement of new roads, widening the road or maintaining the roads

1. Determine road utility pattern

Understand pattern of how different road in region is utilized

1. Provide summarized information

Get summarized information on transportation

During warehousing for transportation department of city council, data were extracted, transformed and loaded into ware house. This itself is an important phase of data mining. Later various reports were generated, meaning data were transformed into information. Which is another important aspect of datamining. Now these reports are required to be thoroughly studied/analyzed to understand trend or pattern. Another technique used in this project is creation of pivot tables and pivot char to generate report as per requirement of city council.

## Data Mining Techniques

There are various data mining techniques available. Some of the key techniques are Association, Clustering, Prediction, Sequential patterns, Decision trees, Combinations, Long-term (memory) processing. For the current project to city council Association techniques can be used. Association technique is most familiar data mining technique. Without going into details of all other techniques, this document briefly discusses association technique and implementation on current scenario.

## Association Technique

Here relationship between two or more items are made to recognize patterns. For example, when a customer buys milk, he also buys sugar. This helps to understand habit of customer. Hence e-commerce site can suggest customer to buy sugar each time they buy milk. Eventually making good impact on sales of sugar. This technique can be useful for transport department as they can identify the pattern of how people uses road. This information can be used while planning strategies, policies or new road. This can help engineers to plan where overhead bridge or underground panel can be more useful. However, other techniques should also be utilized if feasible to get more useful and accurate information.

# Bibliography

* Berry, M. and Linoff, G. 2004. *Data mining techniques*. Indianapolis: Wiley.
* Dunham, MH 2003, *Data Mining: Introductory And Advanced Topics*, n.p.: Upper Saddle River, N.J. : Prentice Hall/Pearson Education, c2003., Leeds Beckett University Library Catalogue, EBSCO*host*, viewed 9 November 2016.
* Han, J, Kamber, M, & Pei, J 2012, *Data Mining. [Electronic Resource] : Concepts And Techniques*, n.p.: Amsterdam ; London : Morgan Kaufmann, c2012., Leeds Beckett University Library Catalogue, EBSCO*host*, viewed 9 November 2016.