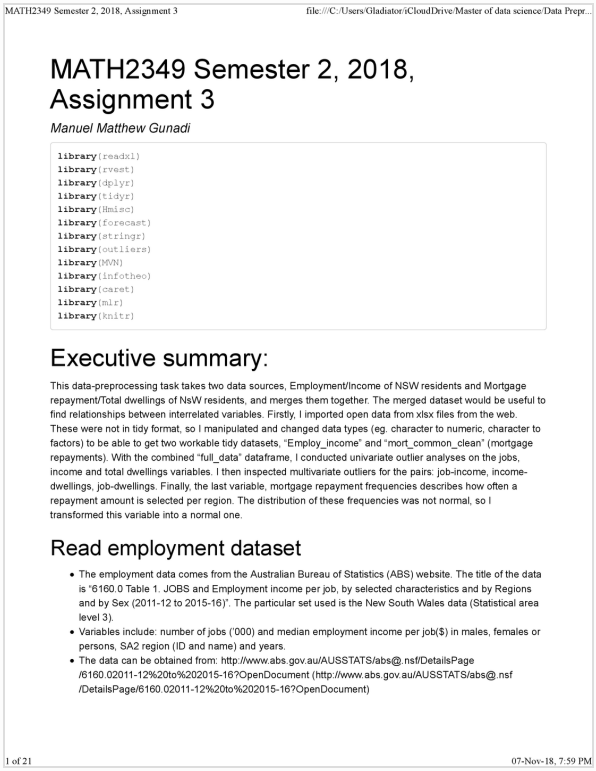
Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Manuel Gunadi
Assignment title:	Assignment 3:
Submission title:	Manuel Gunadi Assignment 3 - MAT.
File name:	s3740473_-_Manuel_Matthew_Gu...
File size:	0
Page count:	21
Word count:	3,865
Character count:	20,308
Submission date:	08-Nov-2018 06:42PM (UTC+1100)
Submission ID:	1035212247



MATH2349 Semester 2, 2018, Assignment 3 file:///C:/Users/Gladiator/iCloudDrive/Master of data science/Data Prepr...

MATH2349 Semester 2, 2018, Assignment 3

Manuel Matthew Gunadi

```
library(readxl)
library(rvest)
library(dplyr)
library(tidyr)
library(Hmisc)
library(forecast)
library(stringr)
library(outliers)
library(MVN)
library(infotheo)
library(caret)
library(mlr)
library(knitr)
```

Executive summary:

This data-preprocessing task takes two data sources, Employment/Income of NSW residents and Mortgage repayment/Total dwellings of NsW residents, and merges them together. The merged dataset would be useful to find relationships between interrelated variables. Firstly, I imported open data from xlsx files from the web. These were not in tidy format, so I manipulated and changed data types (eg. character to numeric, character to factors) to be able to get two workable tidy datasets, "Employ_income" and "mort_common_clean" (mortgage repayments). With the combined "full_data" dataframe, I conducted univariate outlier analyses on the jobs, income and total dwellings variables. I then inspected multivariate outliers for the pairs: job-income, income-dwellings, job-dwellings. Finally, the last variable, mortgage repayment frequencies describes how often a repayment amount is selected per region. The distribution of these frequencies was not normal, so I transformed this variable into a normal one.

Read employment dataset

- The employment data comes from the Australian Bureau of Statistics (ABS) website. The title of the data is '6160.0 Table 1. JOBS and Employment income per job, by selected characteristics and by Regions and by Sex (2011-12 to 2015-16)'. The particular set used is the New South Wales data (Statistical area level 3).
- Variables include: number of jobs ('000) and median employment income per job($) in males, females or persons, SA2 region (ID and name) and years.
- The data can be obtained from: http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/6160.02011-12%20to%202015-16?OpenDocument (http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/6160.02011-12%20to%202015-16?OpenDocument)

1 of 21 07-Nov-18, 7:59 PM