##### INDUSTRIAL TRAINING

**Undergone at**

**JASIEL CREATIONS**

***SECTOR 46, GURGAON***

##### A PRESENTATION REPORT

On

###### **Traffic and Road Safety Analysis (Data Science)**

###### ***Submitted by***

##### MUSKAN SINGHAL

**RA1711008010070**

***in partial fulfillment for the award of the degree***

***of***

##### BACHELOR OF TECHNOLOGY

**IN**

**INFORMATION TECHNOLOGY**

****

SEPTEMBER 2019

**DECLARATION**

I hereby declare that the presentation report submitted titled **“Traffic and Road Safety Analysis (Data Science)”,** is a record of my industrial training programme which I had undergone in the company **Jasiel Creations, sector 46, Gurgaon** during the end of the fourth semester between the period **3rd** **June, 2019** to **25th June, 2019.**

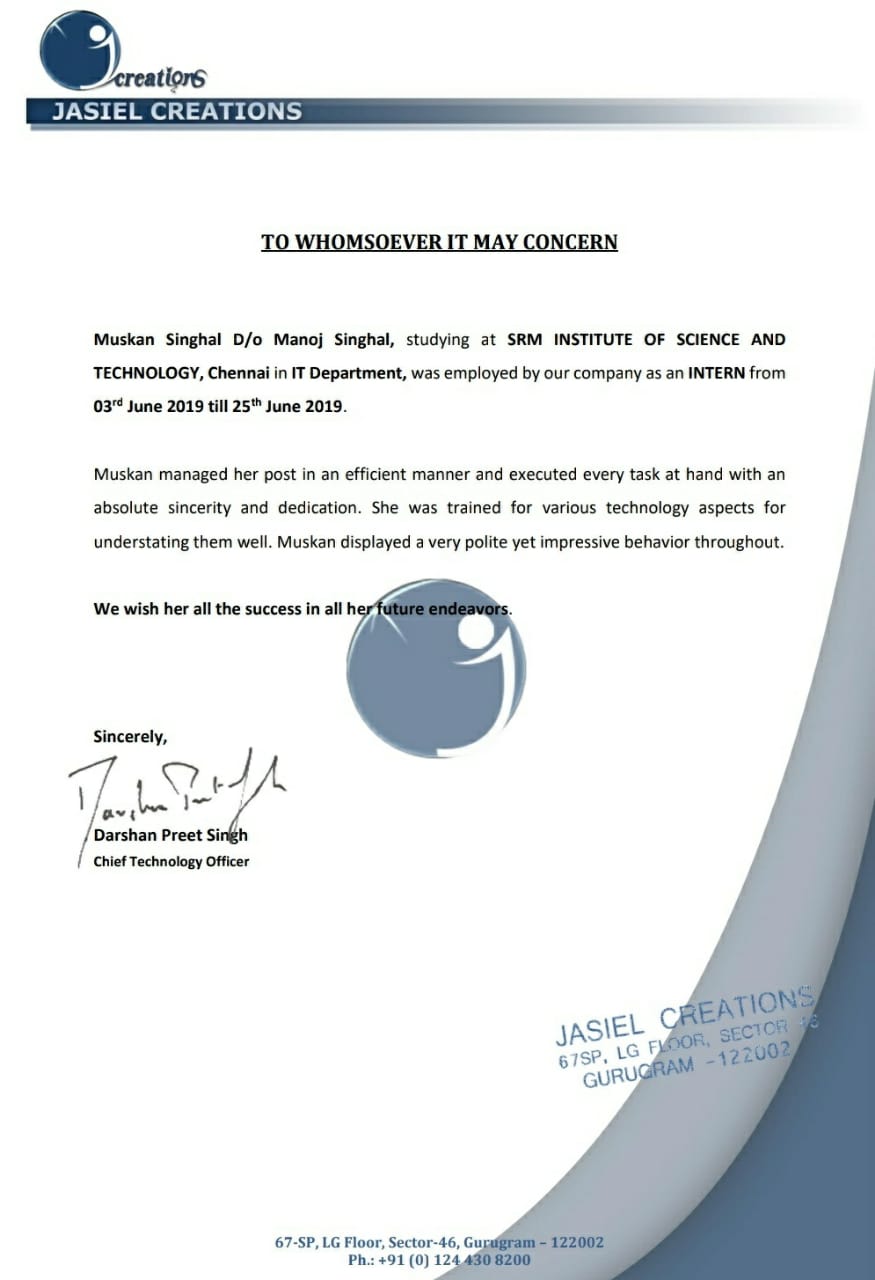
Date : **12th September, 2019**

Name : **Muskan Singhal**

Register Number : **RA1711008010070**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| Serial No. | Topic | Page No. |
| 1. | Industrial Training Certificate | 6 |
| 2. | Company Details | 7 |
| 3. | Introduction to data science | 9 |
| 4. | Applications of data science | 10 |
| 5. | Working on R and Rstudio | 11 |
| 6. | Intro to my research | 12 |
| 7. | Regression with ARIMA errors | 17 |
| 8. | Different road safety progress | 18 |
| 9. | Different road safety problems | 19 |
| 10. | Multi level analysis model | 20 |
| 11. | Time series analysis | 21 |
| 12. | Time series forecasting | 25 |
| 13. | Dickey Fuller test | 29 |
| 14. | Summary | 31 |
| 15. | Conclusion | 32 |

****

* **JASIEL CREATIONS**

Jasiel, founded in 2009, is a global Design and Technology firm specialized in emerging era of interactive communication with many successful executions.

It is headquartered in Melbourne, Australia with two other branches in Gurugram, India and Dubai.

Strategic partnerships with Research & Development houses across the globe enriches them with current and future market trends in the field of technology. Possessing not only the power to innovate, but to customize the innovation have lead them & their customers to endless opportunities with 100% customizable solutions.

With a team of young engineers, entrepreneurs and elite intellectuals, Jasiel Creation’s mission is to revive the industry with continuous novel experience.

Jasiel Creations aim not only at possessing the power to innovate but also to customize the innovation not only in India, Dubai, Australia but in the entire world.

**Core Activities**

They have showcased their work for multiple clients providing products in :

* Virtual Floor
* Interactive Wall
* Multi-Touch Surfaces
* Augmented Reality
* Virtual Mannequin
* Multi-Touch Kiosks
* 3D Depth Sensing Games
* 3D MoCap
* Live Polling
* Augmented Reality with gesture Control
* Instant Photo Booths
* RFID/Bar Code Readers and Scanners
* Android IPad App

They have worked on designing hi-end technology arena where the user has the liberty to play in the virtual world without with ease moments of gestures including (hands, foot taping, head, body, shoulder moments etc).

**Address**

67SP, LG Floor, sector 46

Gurugram, 122002, Haryana, India.

**Website**

jasielcreations.com

**INTRODUCTION TO DATA SCIENCE**

Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management and preservation of large collections of information.

A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product.

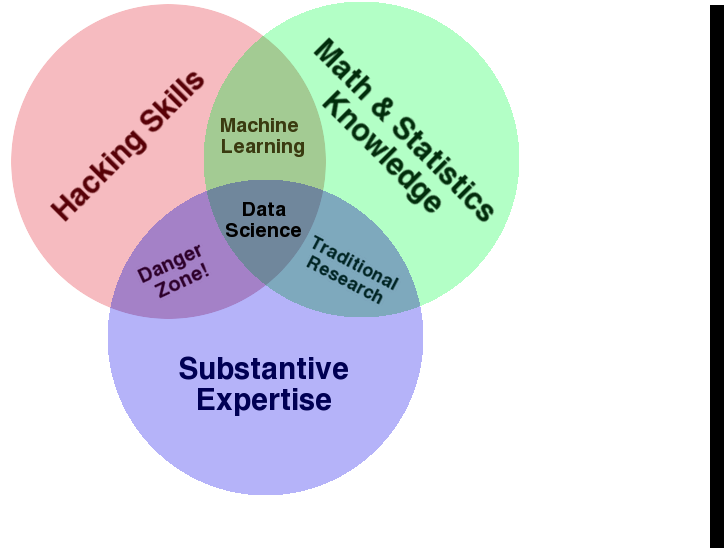
Data scientists tend to be “hard scientists”, particularly physicists, rather than computer science majors. Physicists have a strong mathematical background, computing skills, and come from a discipline in which survival depends on getting the most from the data. They have to think about the big picture, the big problem.

Specifically, Data Science is about the whole processing pipeline to extract information out of data. Data Scientiss understand and care about the whole data pipeline. A data pipeline consists of 3 steps:

1) Preparing to run a model

2) Running the model

3) Communicating the results



**APPLICATIONS OF DATA SCIENCE**

As it is quite clear, Data Science is a broad term, and so are its applications. Almost every application on our smartphone thrives on data. So, it’s only fair to say that it’s practically impossible to list down all the applications of data science because of its sheer omnipresence.

**But we can have a look at some of the broad fields that are using the magic of Data Science:**

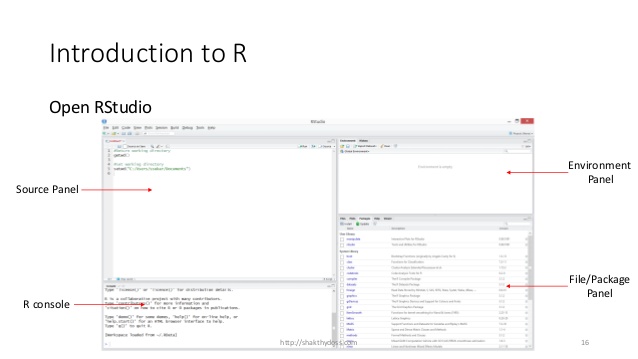
1. Internet Search:  How does Google return such \*accurate\* search results within a fraction of a second? With the use of data science!
2. Recommendation Systems:  From “people you may know” on Facebook or LinkedIn to “people who’ve bought this product also liked…” on Amazon to your daily curated playlists on Spotify to even “suggested videos” on YouTube, everything is fueled by Data Science.
3. Image/Speech/Character Recognition:  This pretty much goes without saying. What do you think is the brain behind “Siri”, if not Data Science? Also, how do you think Facebook recognizes your friend when you upload a photo with them? It’s not magic; it’s science – Data Science.
4. Gaming:  EA Sports, Sony, Nintendo, Zynga, and other giants in this domain have taken it upon themselves to take your gaming experience to an altogether new level. Games are now developed and improved using Machine Learning algorithms so that it can upgrade as you move up to higher levels.
5. Price Comparison Websites:  These websites are fuelled by data. For them, the more the merrier. The data is fetched from the relevant websites using APIs. PriceGrabber, PriceRunner, Junglee, Shopzilla are some such websites.

**WORKING ON R and Rstudio**

R is the name of the programming language used itself and RStudio is a convenient interface. People all over the world are turning to R, an open source statistical language, to make sense of data. Inspired by the innovations of R users in science, education, and industry, RStudio develops free and open tools for R and enterprise ready professional products for teams to scale and share work.

Rstudio is encouraged to be used as :

1. Provides community support.
2. Enforces best practices like documentation, version control, unit testing.
3. Integrates R packages, documentation, debugging and viewing tools.

****

The panel in the upper right contains your workspace as well as a history of the commands that you’ve previously entered.

Any plots that you generate will show up in the panel in the lower right corner. The panel on the left is where the action happens. It’s called the console. Everytime you launch RStudio, it will have the same text at the top of the console telling you the version of R that you’re running.

Below that information is the prompt. As its name suggests, this prompt is really a request, a request for a command.

Initially, interacting with R is all about typing commands and interpreting the output. These commands and their syntax have evolved over decades (literally) and now provide what many use.

**INTRODUCTION TO MY RESEARCH**

In this thesis we present a comprehensive study into novel series models for aggregated road safety data. The models are mainly intended for analysis of indicators relevant to road safety, with a particular focus on how to measure these factors. Such developments may need to be related to or explained by external inﬂuences. It is also possible to make forecasts using the models. Relevant indicators include the number of persons killed per month or year. These statistics are closely watched by government agencies and the public, and their relevance to society is not disputed. A large body of research is devoted to the improvement of road safety. To that end, changes in the number of accidents or victims are often attempted to be explained by factors such as exposure, policy, driving under the inﬂuence of alcohol, speeding by drivers. Some factors such as policy changes can be directly observed.

All these tests and analysis are performed and experimented on using multiple data science related techniques and series such as

* Descriptive and qualitative analysis
* Linear regression
* Multilevel analysis
* Time series analysis
* Cost-Benefit Analysis

- Univariate and multivariate time series model

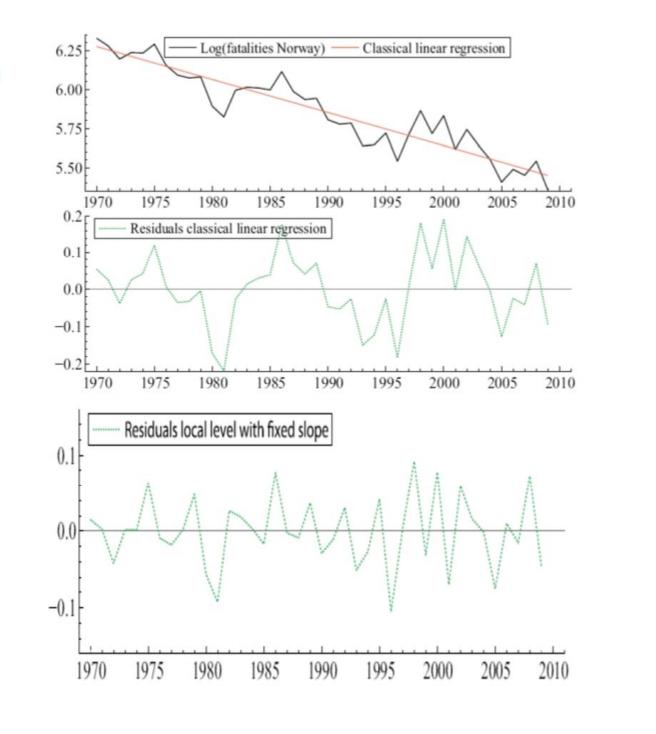
Qualitative and descriptive research methods have been very common procedures for conducting research in many disciplines, including education, psychology and social sciences.

The terms qualitative and descriptive research are sometimes used inter-changeably. However, a distinction can be made between them. One fundamental characteristics of both types of research is that they involve naturalistic data. The major goal is to describe a phenomenon and its characteristics.

Linear regression models are used to show or predict the relationship between two variables or factors. The factor that s being predicted or for which the equation is solved for is called the dependent variable. The factor used to predict the value of the dependent variable is called as the independent variables.

Regression analysis is commonly used in research as it establishes that a correlation exists between variables. But that is not same as causation. Even a line in a simple linear regression that fits the data points well may not say something definitive about a cause and effect relationship.

Current research models encounter a large number of false positives and with changing characteristics of the time series, these models require additional training. Time series analysis is done to predict the future values of the series using current information from the dataset.



Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, astronomy. In short, almost any domain which involves temporal measurements.

Most time series patterns can be described in terms of two basic classes of components:

* Trend
* Seasonality

Trend represents a general systematic linear or nonlinear component that changes over time and does not repeat within the time range captured by the data whereas, seasonality has formally similar nature however, it repeats itself in systematic intervals over time. Trend and seasonality can co-exist too.

The components of the time series are as complex and sophisticated as the data itself. With increasing time, the data obtained increases and it doesn’t always mean that more data means more information but, larger sample avoids the error that arises due to random sampling.

### The term “univariate time series” refers to a time series that consists of single or scalar observations recorded sequentially over equal time increments. Some examples are monthly CO2 concentrations and southern oscillations to predict el nino effects.

Whereas multivariate time series models are designed to capture the dynamic of multiple time series simultaneously and leverage dependencies across these series for more reliable predictions.

In the case of predicting the temperature of a room every second univariate analysis is preferred since there is only one unit that is changing.  But to calculate the altitude of the rocket from the time of its launch, a multivariate time series analysis comes in handy as there will be other changes like reduction in fuel with time.

In the case of economics, multivariate time series are used to understand how policy changes to one variable, for example, an interest rate, may affect other variables over different horizons.

The data ingested for analysis comes with a lot of non-linearities and these fluctuations have to be smoothed out to make sense out of the data.

Usually, time series models are  [adequately approximated](http://www.statsoft.com/Textbook/Time-Series-Analysis) by a linear function; if there is a clear monotonous nonlinear component, the data first need to be transformed to remove the nonlinearity. Usually, logarithmic, exponential, or polynomial function are used.

During the 28 days course of my internship at Jasiels, we performed different types of tests on two topics, majorly, **Traffic and Road Safety Analysis,** and **Weather Analysis and Forecasting.**

For the researching purposes, we were taught and made to work upon the different models in linear Time Series which included –

* Autoregressive model (AR)
* Moving Average model (MA)
* Auto Regressive-Moving-Average model (ARMA)
* Auto Regressive Integrated Moving-Average model (ARIMA)

In statistics, econometrics and signal processing and an **autoregressive** (**AR**) **model** is a representation of a type of random processes as such, it is used to describe certain time-varying processes in nature, economics etc. The autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic term (an imperfectly predictable term), thus the model is in the form of a stochastic difference equation. It is also a special case of the **vector autoregressive model** (VAR), which consists of a system of more than one interlocking stochastic difference equation in more than one evolving random variable

The **moving-average model** (**MA model**), also known as **moving-average process**, is a common approach for modelling univariate time series. The moving-average model specifies that the output variable depends linearly on the current and various past values of a stochastic term.

Together with the **MA model**, **AR model** is a special case and key component of the more general **ARMA** and **ARIMA** models of time series, which have a more complicated stochastic structure.

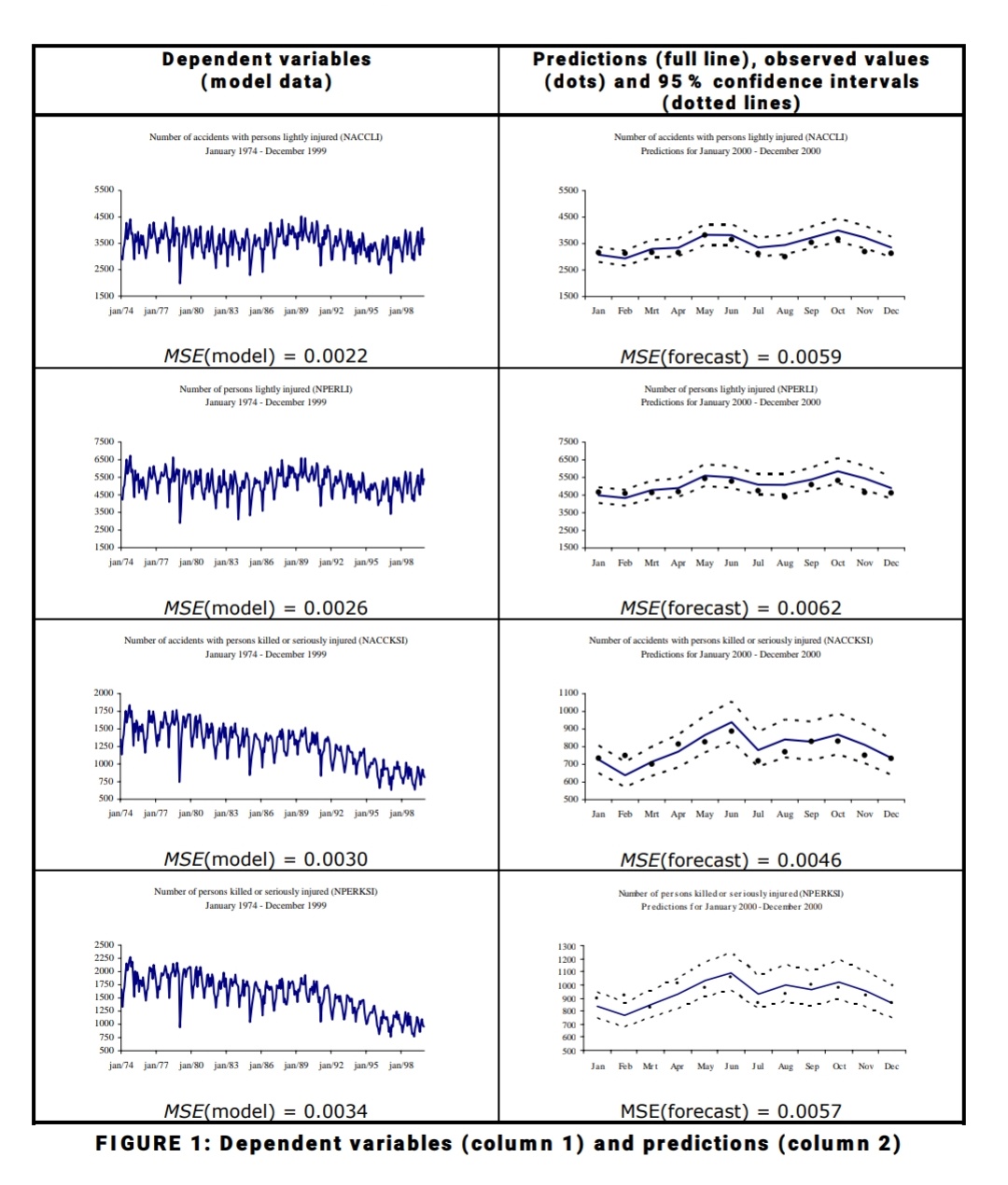
Contrary to the moving-average model, the autoregressive model is not always stationary as it may contain a unit root.

In the analysis of time series, **auto regressive moving average models** provide a parsimonious description of a weakly stationary stochastic process in terms of two polynomials, one for the autoregression (AR) and the second for the moving average (MA).

Given a time series of data *Xt* , the ARMA model is a tool for understanding and, perhaps, predicting future values in this series. The AR part involves regressing the variable on its own lagged (i.e., past) values. The MA part involves modeling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past. The model is usually referred to as the ARMA(*p*,*q*) model where *p* is the order of the AR part and *q* is the order of the MA part.

The **autoregressive integrated moving average (ARIMA)** model is a generalization of the ARMA model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series. ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step can be applied one or more times to eliminate the non-stationarity.

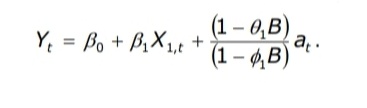
The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The I indicate that the data values have been replaced with the difference between their values and the previous values. The purpose of each of these features is to make the model fit the data as well as possible.



In this study, dependent traffic safety variables are expressed in terms of independent explanatory variables. Multiple linear regression can be used to model a relationship between a dependent variable and one or more independent variables. It allows investigating the effect of changes in the various factors on the dependent variable. If the observations are measured over time, the model is called a time series regression. The resulting statistical relationship can be used to predict future values of the target. If one is interested in the explanatory and predictive power of the regression equation, all necessary assumptions should be met. To this end, a regression model with ARIMA errors will be used as a means to analyze traffic accident time series data.

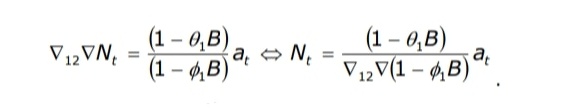
**REGRESSION WITH ARIMA ERRORS**

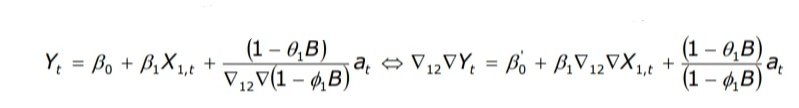
The ARIMA modeling approach can now be applied to the multiple regression equation to model the information that remains in the error terms. Assume a regression model with one explanatory variable, denoted as Yt = β0 + β1X1,t + Nt. Suppose further that the error terms are autocorrelated, and that they can be appropriately described by an ARMA(1,1) process. This model can then be written as: Yt = β0 + β1X1,t + Nt, where (1–φ1B)Nt = (1– θ1B)at, and at is assumed to be white noise. Substituting the correction for the error term into the regression equation gives:



Because of the specific form in the error terms, the classical least squares methods are not appropriate to estimate the parameters of this equation. Instead, the SAS-ARIMA procedure with Maximum Likelihood estimation is used to set up the models. The Likelihood function is maximized using Marquardt’s method via nonlinear least squares estimation.

If differencing is applied to the errors in a multiple regression, Pankratz shows that all corresponding series (both of the dependent and the explanatory variables) should be differenced. This can be seen from our small regression example. Differencing the error terms twice results in the following expression, with the ARMA(1,1) model now in the differenced error terms:



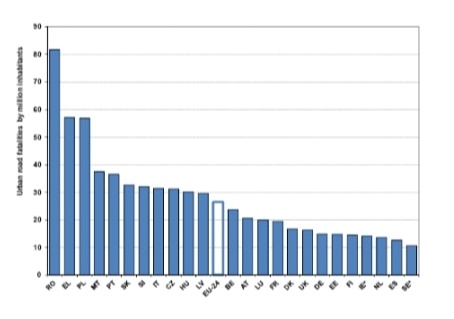
Substituting back this expression into the regression equation gives: 

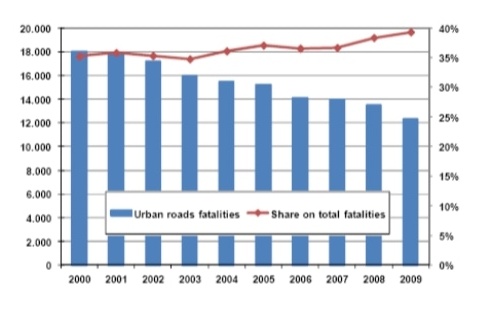
The intercept is now possibly different, but the (theoretical) regression coefficient β1 is not affected by the differencing operation. Its estimated value may differ slightly, since the estimation is done on different (although related) time series.

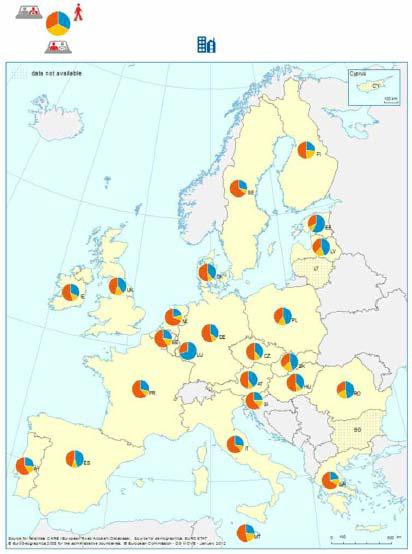
**DIFFERENT ROAD SAFETY PROGRESS IN DIFFERENT COUNTRIES**

The following table and bar graph shows the Road Fatalities changes in urban areas, during 2001-2010, obtained using by working and researching upon different set of values using Rstudio.







**DIFFERENT ROAD SAFETY**

**PROBLEMS IN DIFFERENT**

**COUNTRIES**

The following table results show the percentage of

different road safety problems in different countries in

in suburban and total area through different modes of

transportations that are two-wheelers, cycles,

pedestrians.

These calculations are done to focus on dynamic

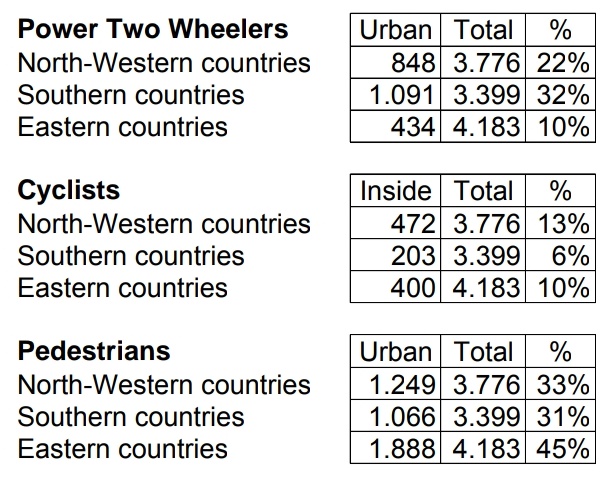
analysis and management of real-time traffic.

Analysis of data traffic and road accidents are

further performed using big data techniques

and tests, social networks data and algorithms

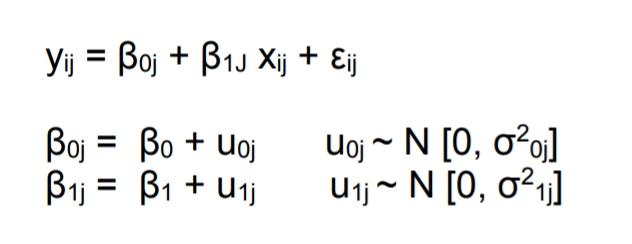
that use data from multiple resources.



**MULTI-LEVEL ANALYSIS MODEL**

Basic multi-level model formulation:

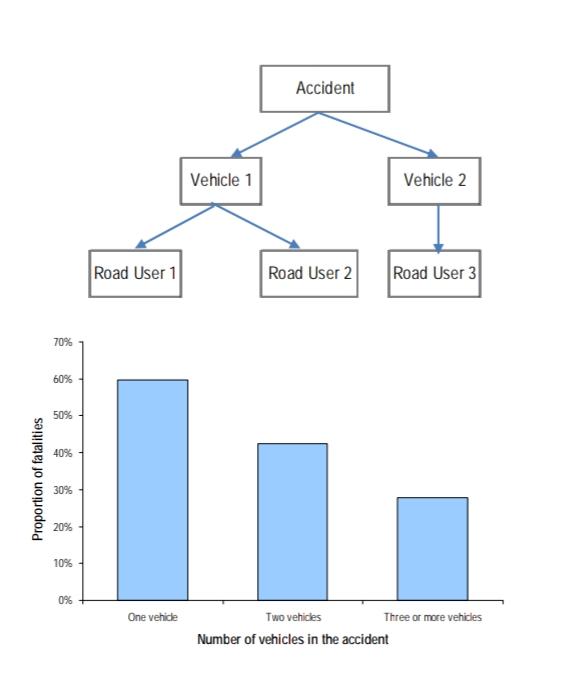
Observations i are nested within units j



ML model formulations include :

* Allow improving the fit of the model to the data,
* Allow identifying and explaining random variation at specific levels of hierarchy considered,
* Yield different conclusions than single-level model formulations with respect to the significance of the parameter estimates.

**Testing on risk and protection factors in fatal accidents**

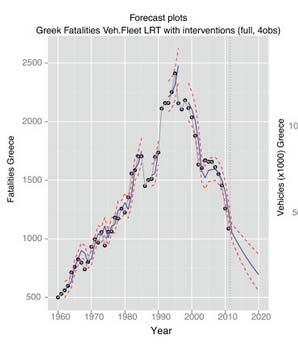


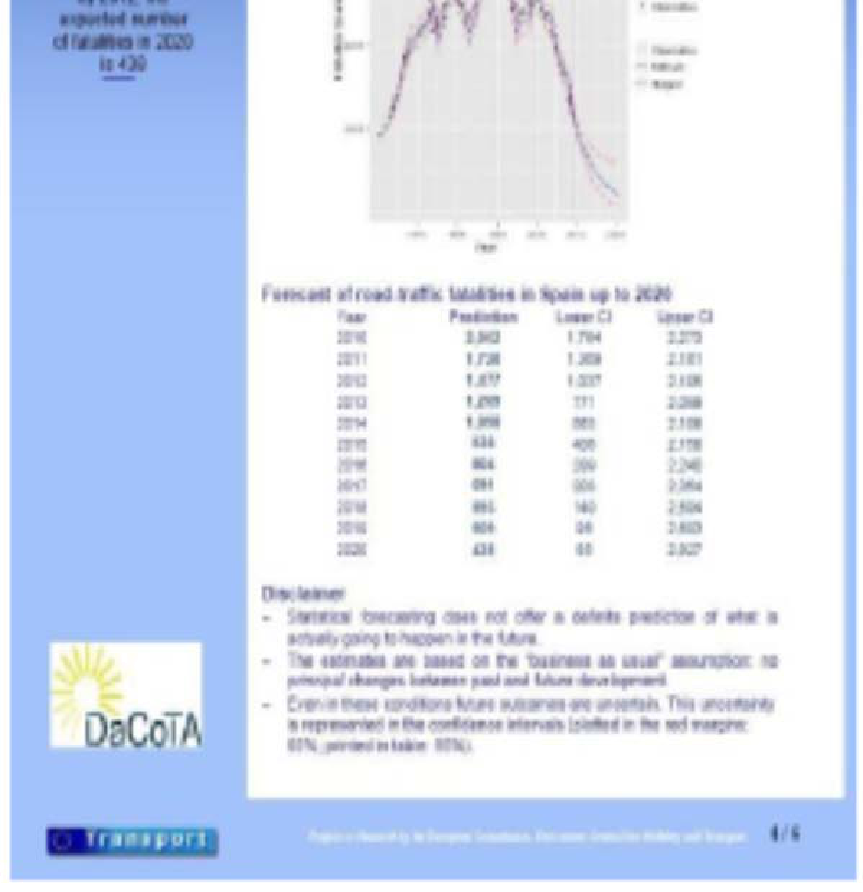
* The above diagram shows the obtained accident hierarchy multi-level model
* Disaggregate data for 1000 fatal accidents from in-depth investigation methods for 7 European countries.
* The fatality risk largely depends on the vulnerability of the road user compared to the vulnerability of the collision opponent.
* Several road, traffic and individual factors interact in terms of injury severity.
* The ‘vehicle’ random effect is necessary to better capture these effects.

**TIME SERIES ANALYSIS**

Road safety observations are often serially correlated – autocorrelation.

As we have already discussed about the types and attributes of time series analysis, we will move forward to the testing of ‘**forecasting fatalities in European countries and in Greece’.**





**Figure a Figure b**

* Data for the period 1960 to 2010 were used to forecast fatalities 2011-2020 in 30 European countries (figure a) and in Greece (figure b) by means of state- space models.

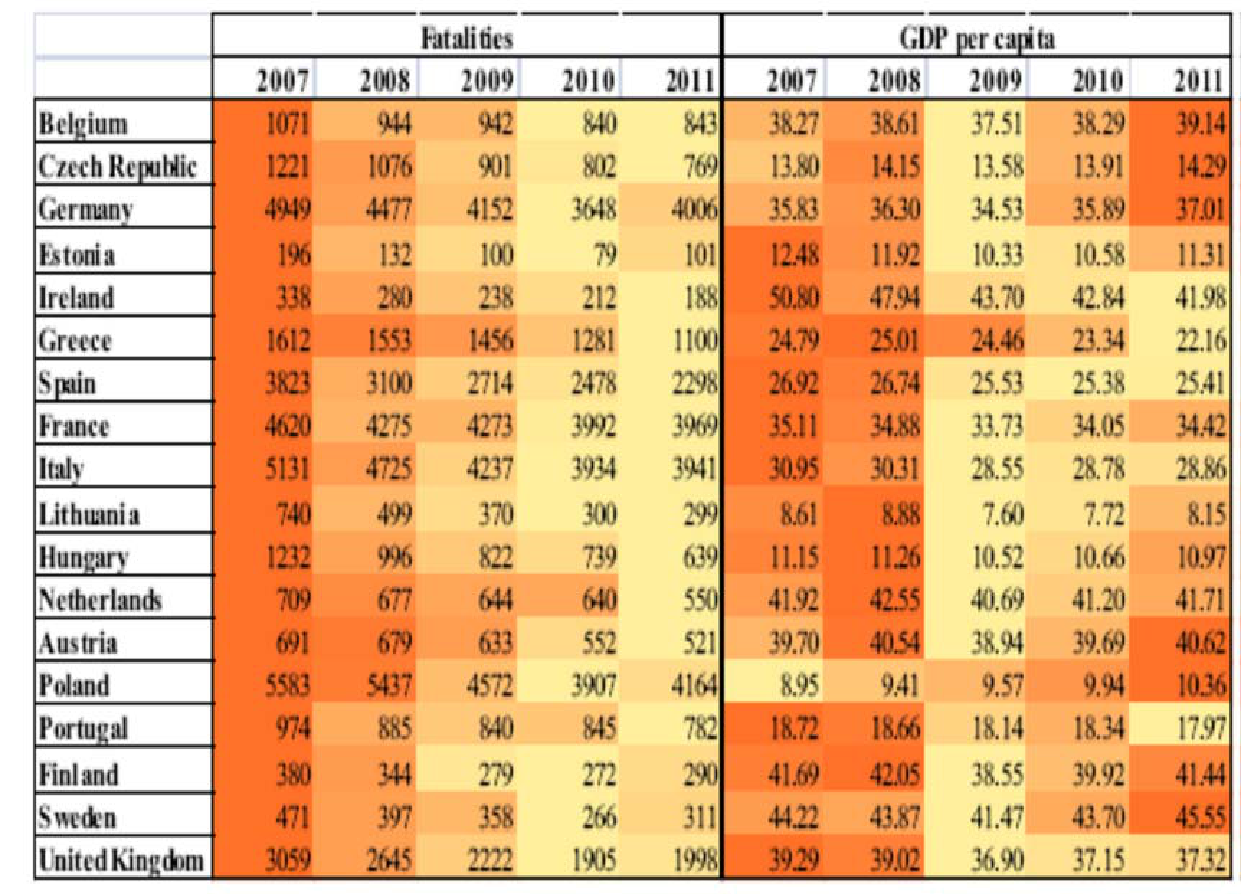
For figure a,

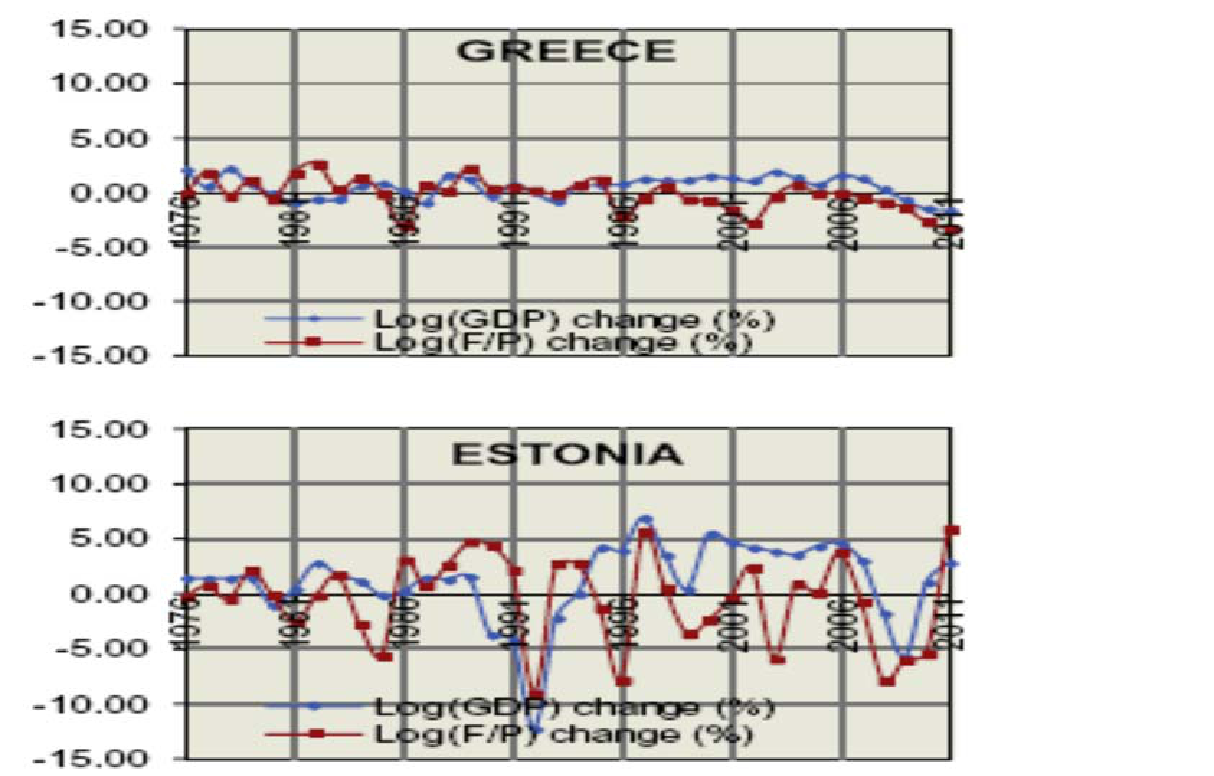
* Explicitly examined the presence or not of correlation between fatality series and exposure series.
* If fatalities and exposure are correlated, then the forecasts can be made on the basis of mobility scenarios like 20% increase in mobility by 2020.

For figure b,

* Exposure and fatalities are found to be related
* Intervention variables are introduced to capture shocks/breaks in the series
* An intervention variable to model the current economic recession, assuming it ends on 2013.
* Data for the period 2009-2011 are used to validate the model.
* Forecasts for 2012-2020: actual data for 2012, 2013 lie within the 955 confidence intervals of the forecasts.

Another testing done is ‘**effects of GDP changes on road fatalities’**





* The economic recession appears to have affected road fatalities in all European countries, due to decrease in mobility and other factors.
* Modelling the annual change of fatality rate against annual change on GDP.
* Statistically significant relationship between annual GDP increase and fatality rate increase was established, as well as a statistically significant relationship between annual GDP decrease and fatality rate decrease.
* Particularly in northern/western European countries, annual GDP decrease is associated with fatality rate decrease in the same year, as well as on one year later.
* Once the socioeconomic conditions improve, fatalities may temporarily increase.

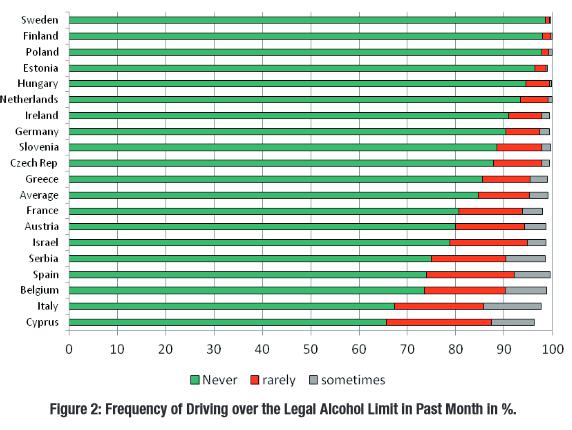
**UNDERSTANDING ROAD USER BEHAVIOUR**

Human factors are the basic causes of road accident in 65-95% of road accidents.

Human factors include a large number of specific factors that may be considered as accident causes including:-

1. Driver injudicious actions including speeding and traffic violations.
2. Driver error reaction like loss of control, failure to keep safe distances, sudden braking etc.
3. Behavior or inexperience like aggressive driving, nervousness, uncertainty etc.
4. Driver distraction or impairment like alcohol, fatigue, use of mobile phone etc.

**MONITORING ROAD SAFETY ATTITUDES AND PERCEPTIONS**

The given diagram shows the SARTRE survey

i.e. Social Attitudes to Road Traffic Risk

in Europe during years 1993, 1996, 2003, 2012.

A common questionnaire on national random

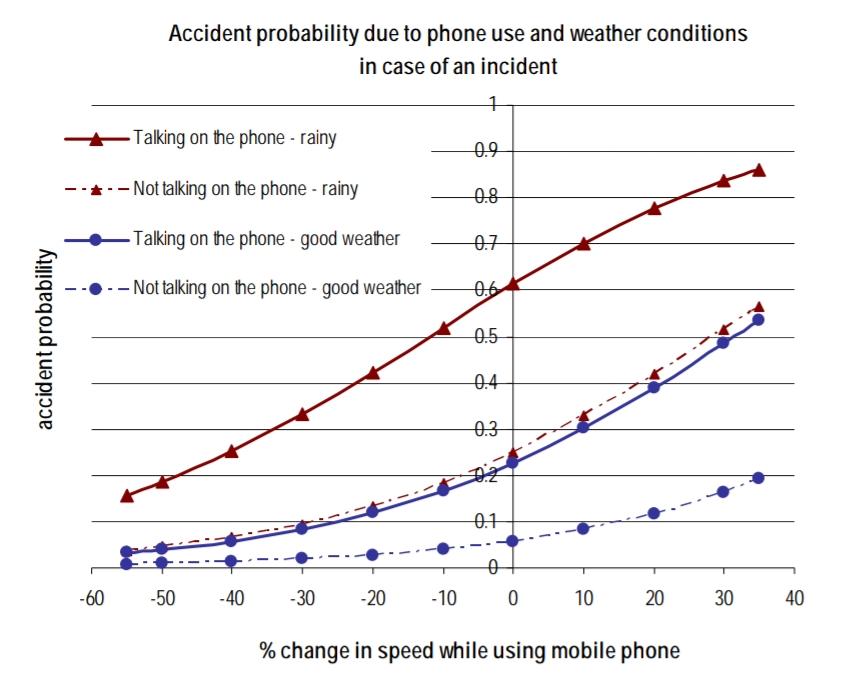
samples of around 1000 participants.

The questionnaire included topics like:-

* mobility and travel motivations
* risk perception
* attitudes towards measures
* driver self-assessment
* self-reported behavior
* speeding, alcohol

**USE OF MOBILE PHONES, DRIVER SPEED AND ACCIDENT PROBABILITY**

* Investigation of the interrelation between mobile phone use, driver speed and accident probability.
* The research focuses on the behavior of 30 young drivers aged between 18 to 30 years old.
* A driving simulator experiment took place, in which participants drove in urban areas, good as well as rainy weather conditions, with and without the use of mobile phones.
* Binary logistic regression methods were used to analyze the combined influence of mobile phone, driver speed and other parameters on the probability of an accident.



The study shows that use of mobile phone while driving leads to increase in accident probability and significant decrease of mean speed in urban and interurban environment.

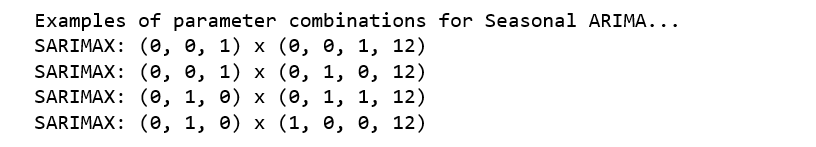
**TIME SERIES FORECASTING WITH ARIMA**

We are going to apply one of the most commonly used method for time-series forecasting, known as ARIMA, which stands for Autoregressive Integrated Moving Average.

ARIMA models are denoted with the notation ARIMA(p, d, q). These three parameters account for seasonality, trend, and noise in data:

p = d = q = range(0, 2)  
pdq = list(itertools.product(p, d, q))  
seasonal\_pdq = [(x[0], x[1], x[2], 12) for x in list(itertools.product(p, d, q))]

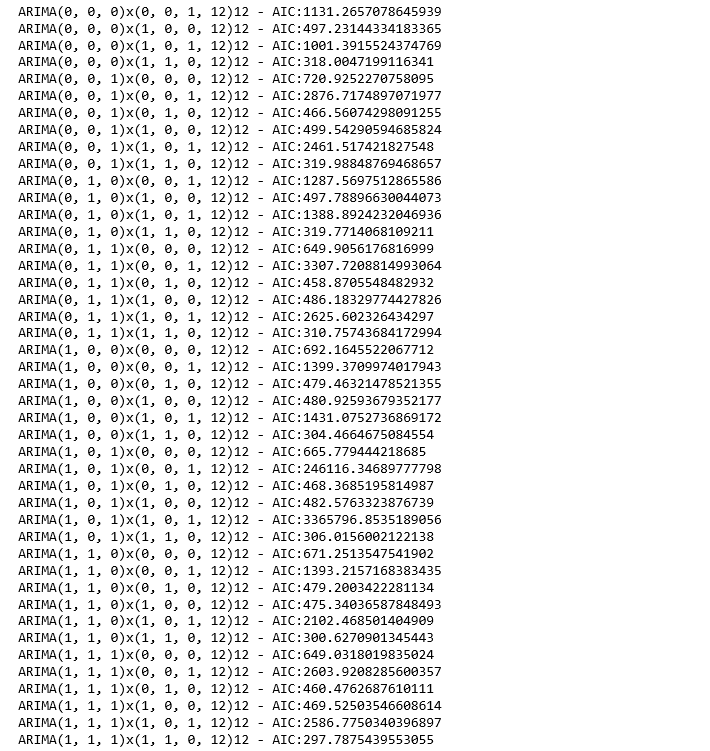
print('Examples of parameter combinations for Seasonal ARIMA...')  
print('SARIMAX: {} x {}'.format(pdq[1], seasonal\_pdq[1]))  
print('SARIMAX: {} x {}'.format(pdq[1], seasonal\_pdq[2]))  
print('SARIMAX: {} x {}'.format(pdq[2], seasonal\_pdq[3]))  
print('SARIMAX: {} x {}'.format(pdq[2], seasonal\_pdq[4]))



This step is parameter Selection for our furniture’s sales ARIMA Time Series Model. Our goal here is to use a “grid search” to find the optimal set of parameters that yields the best performance for our model.

for param in pdq:

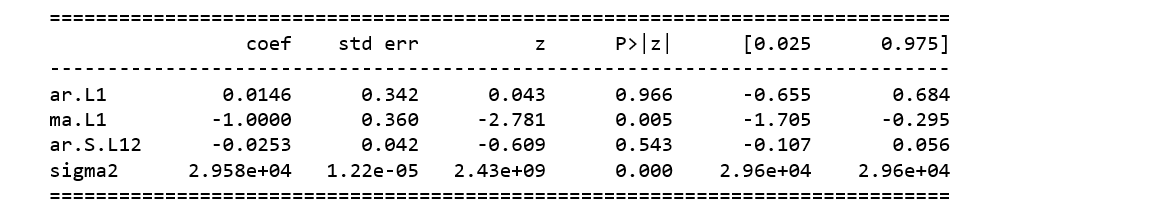
for param\_seasonal in seasonal\_pdq:  
 try:  
 mod = sm.tsa.statespace.SARIMAX(y,  
 order=param,  
 seasonal\_order=param\_seasonal,  
 enforce\_stationarity=False,  
 enforce\_invertibility=False)results = mod.fit()print('ARIMA{}x{}12 - AIC:{}'.format(param, param\_seasonal, results.aic))  
 except:  
 continue



The above output suggests that SARIMAX(1, 1, 1)x(1, 1, 0, 12) yields the lowest AIC value of 297.78. Therefore we should consider this to be optimal option.

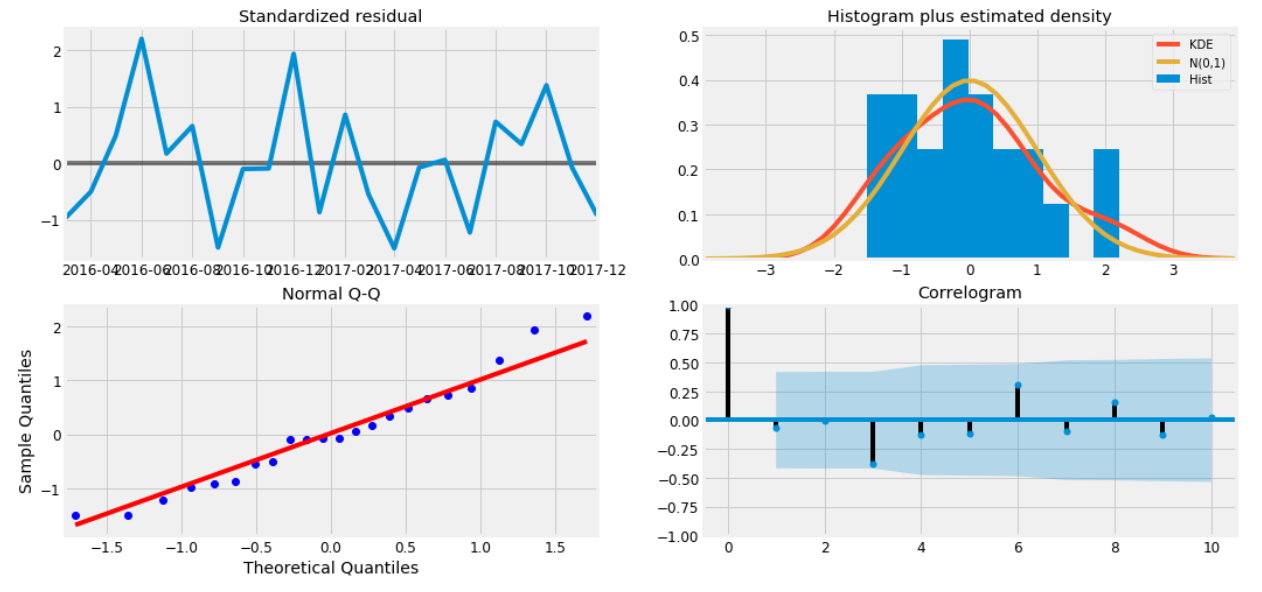
**Fitting the arima model –**

mod = sm.tsa.statespace.SARIMAX(y,  
 order=(1, 1, 1),  
 seasonal\_order=(1, 1, 0, 12),  
 enforce\_stationarity=False,  
 enforce\_invertibility=False)results = mod.fit()print(results.summary().tables[1])



We should always run model diagnostics to investigate any unusual behavior.

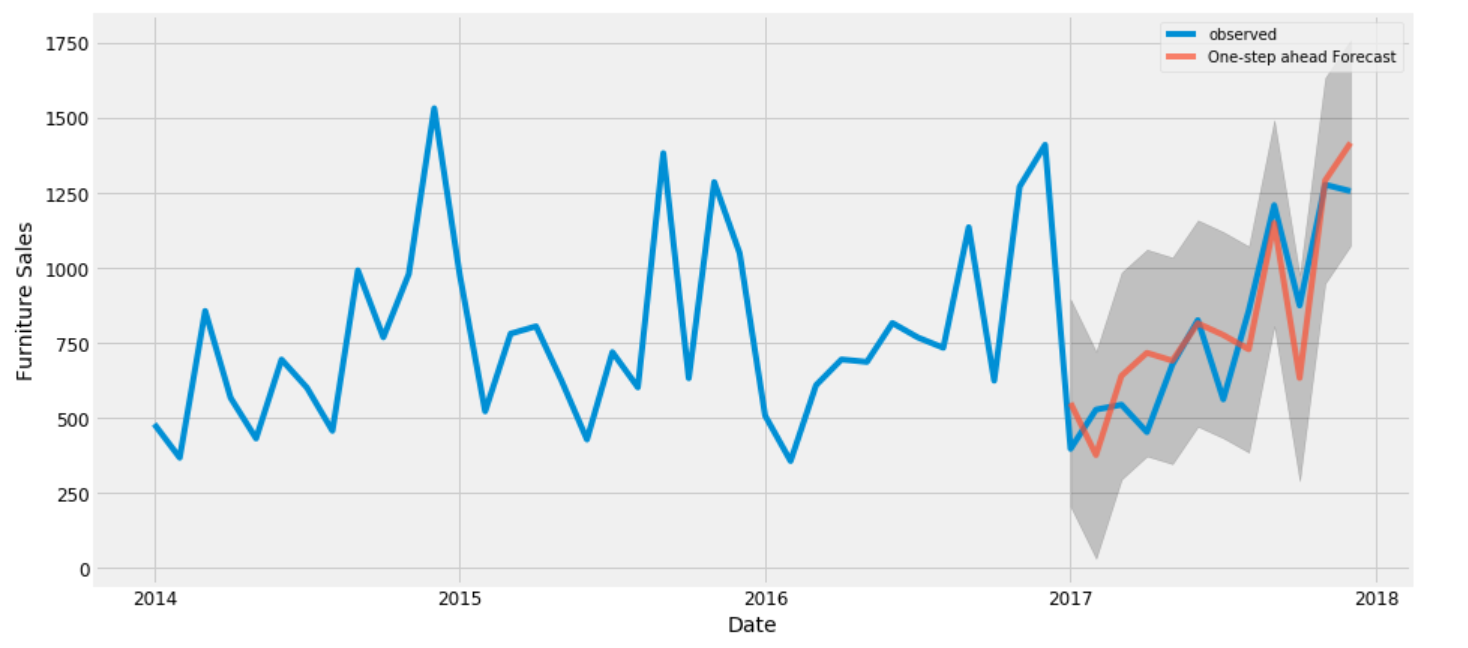
results.plot\_diagnostics(figsize=(16, 8))  
plt.show()



**Validating Forecasts**

To help us understand the accuracy of our forecasts, we compare predicted sales to real sales of the time series, and we set forecasts to start at 2017–01–01 to the end of the data.

pred = results.get\_prediction(start=pd.to\_datetime('2017-01-01'), dynamic=False)  
pred\_ci = pred.conf\_int()ax = y['2014':].plot(label='observed')  
pred.predicted\_mean.plot(ax=ax, label='One-step ahead Forecast', alpha=.7, figsize=(14, 7))ax.fill\_between(pred\_ci.index,  
 pred\_ci.iloc[:, 0],  
 pred\_ci.iloc[:, 1], color='k', alpha=.2)ax.set\_xlabel('Date')  
ax.set\_ylabel('Furniture Sales')  
plt.legend()plt.show()



The line plot is showing the observed values compared to the rolling forecast predictions. Overall, our forecasts align with the true values very well, showing an upward trend starts from the beginning of the year and captured the seasonality toward the end of the year.

y\_forecasted = pred.predicted\_mean  
y\_truth = y['2017-01-01':]mse = ((y\_forecasted - y\_truth) \*\* 2).mean()

print('The Mean Squared Error of our forecasts is {}'.format(round(mse, 2)))

The Mean Squared Error of our forecasts is 22993.58

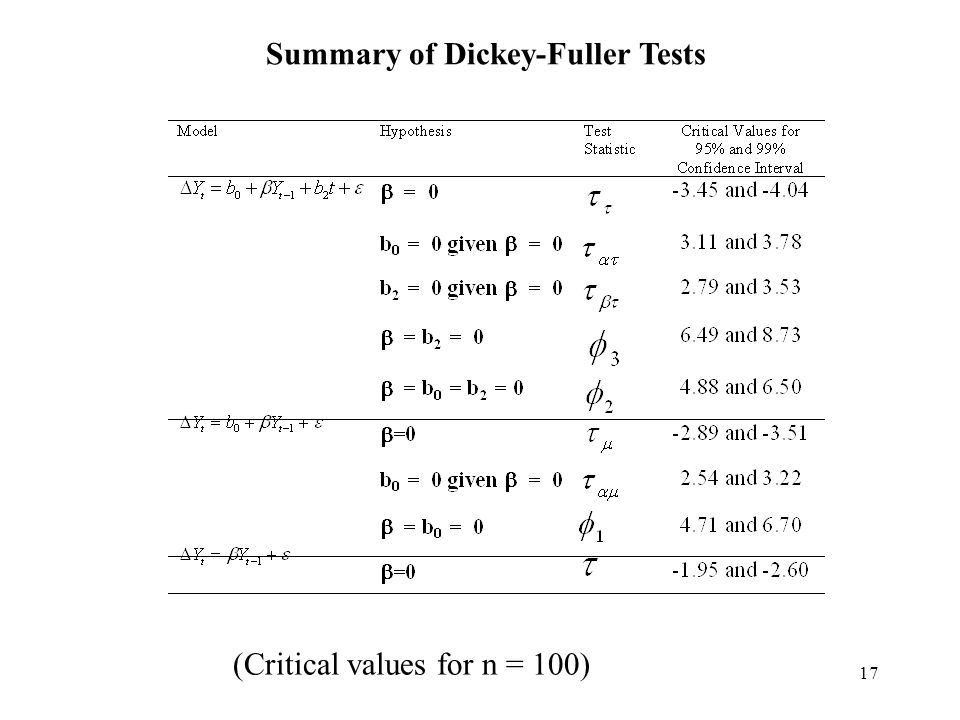
print('The Root Mean Squared Error of our forecasts is {}'.format(round(np.sqrt(mse), 2)))

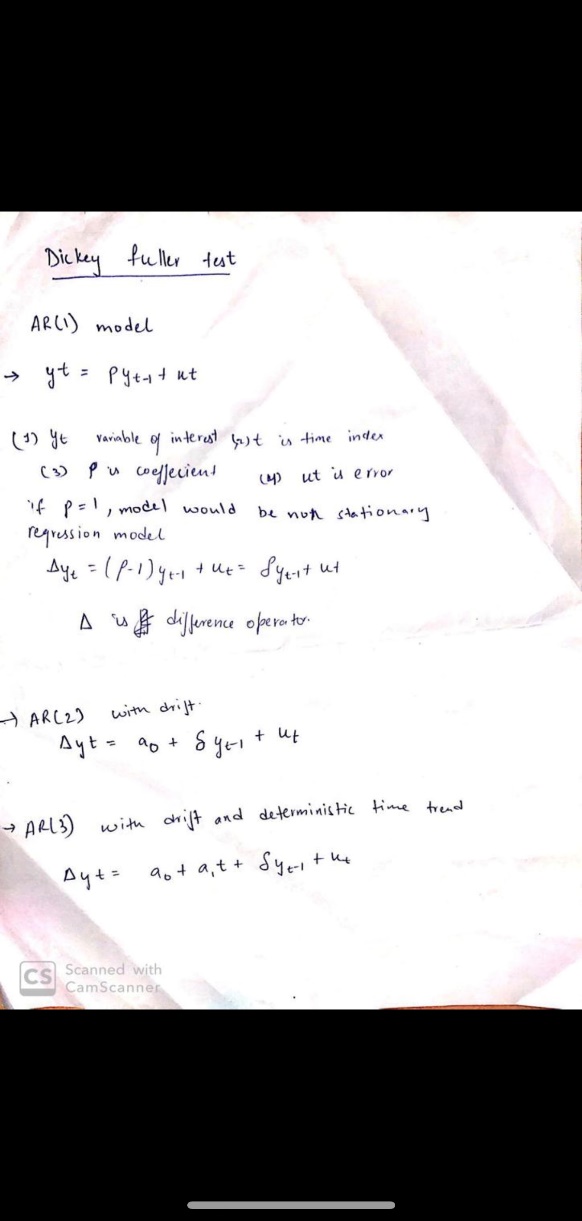
The Root Mean Squared Error of our forecasts is 151.64

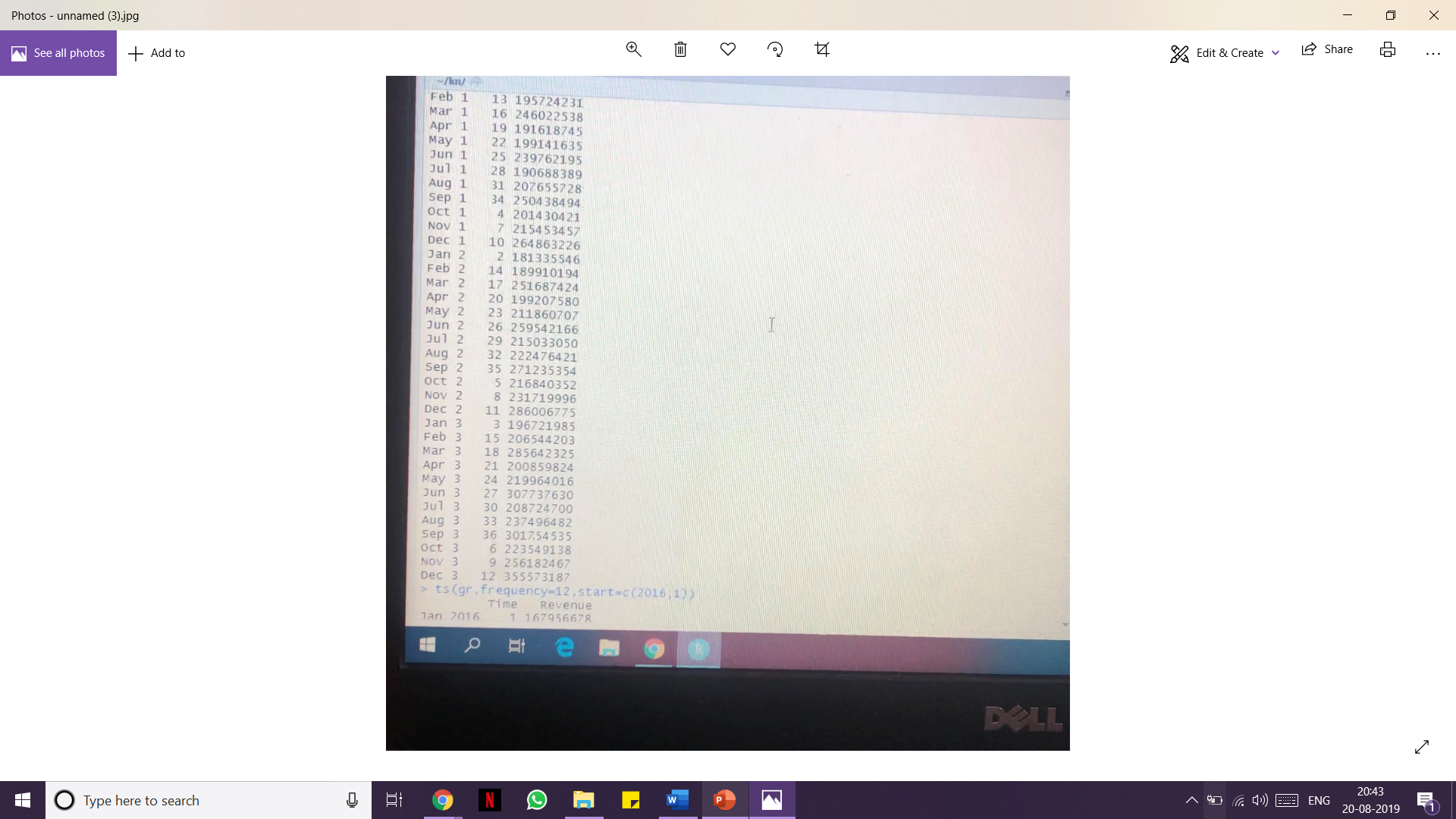
**AUGMENTED DICKEY-FULLER TEST**

In statistics, the **Dickey–Fuller test** tests the null hypothesis that a unit root is present in an autoregressive model. The alternate hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity.

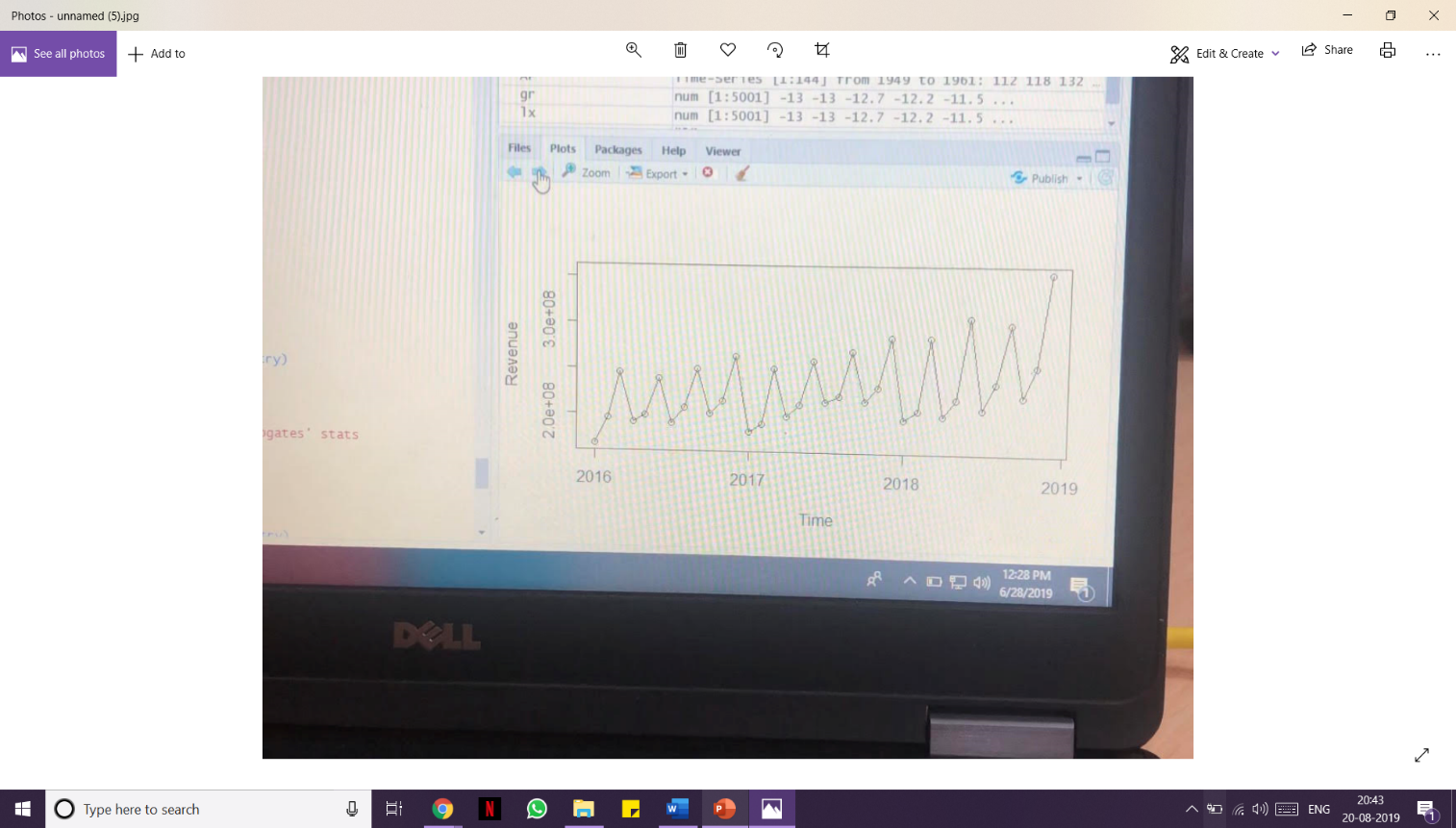
It is named after the statisticians David Dickey and Wayne Fuller, who developed the test in 1979.

{\displaystyle \Delta y\_{t}=a\_{0}+a\_{1}t+\delta y\_{t-1}+u\_{t}\,}

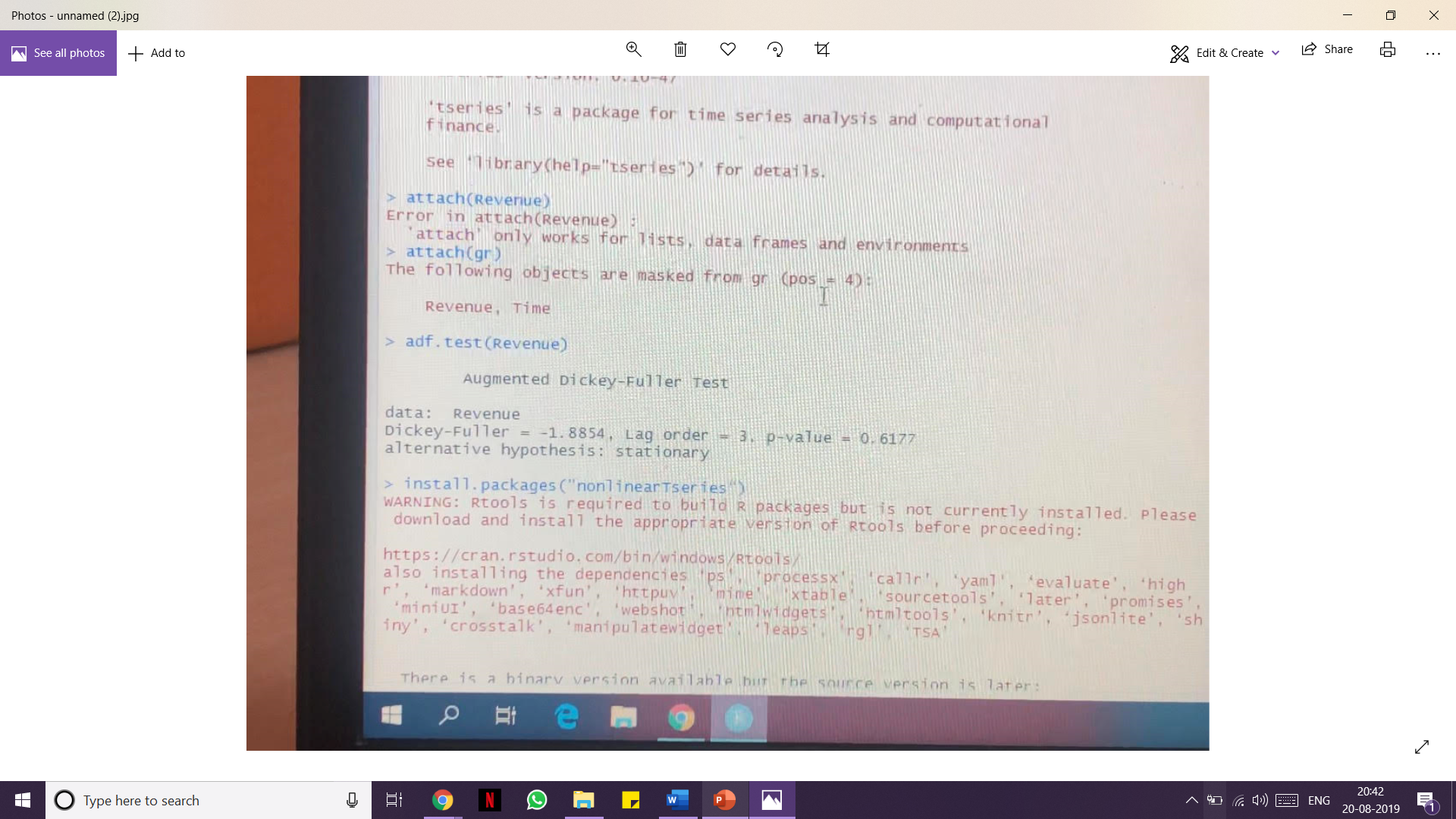
****

****

given set of data

****

graph plotted

****

**ADF** test performed

**SUMMARY**

Despite the important improvements in road safety, the number of road accident casualties around the world is still unacceptable and there is need for intensification of efforts for further improvement.

During my research period, knowledge was gained about different series analysis models and methods specifically

* Descriptive and qualitative analysis
* Linear regression
* Multilevel analysis
* Time series analysis
* Cost-Benefit Analysis

- Univariate and multivariate time series model

where we worked majorly on multi-level analysis and time series analysis.

Current research models encounter a large number of false positives and with changing characteristics of the time series, these models require additional training. Time series analysis is done to predict the future values of the series using current information from the dataset.

Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, astronomy. In short, almost any domain which involves temporal measurements.

**CONCLUSION**

I would like to thank SRM Institute of Science and Technology for providing me with an opportunity to work with a company as an intern and learn through experience on hands.

This opportunity gave me plenty of good lessons based on ethical as well as professional views. I was able to widen my knowledge about Cyber Security and Java Swing. I would also like to thank my teachers and faculties for their constant support and guidance. It was an enriching experience. I am really looking forward to such opportunities to widen my horizon.