WIKIPEDIA · APR. 30, 2020

WIKIPEDIA · APR. 30, 2020

WIKIPEDIA · APR. 30, 2020

# Web Traffic Time-series Forecast

For Wikipedia webpages

By: Sara Alsoghayer

GENERAL ASSEMBLY

# Presentation Agenda

## Focus areas

- Problem statement
- Basic EDA
- Data Modeling
- Predictions
- Challenges
- Future work

# Problem Statement:

Forecasting the future web traffic (webpage visits) of multiple time-series for Wikipedia webpages and analyzing the pattern of traffic.

# Dataset

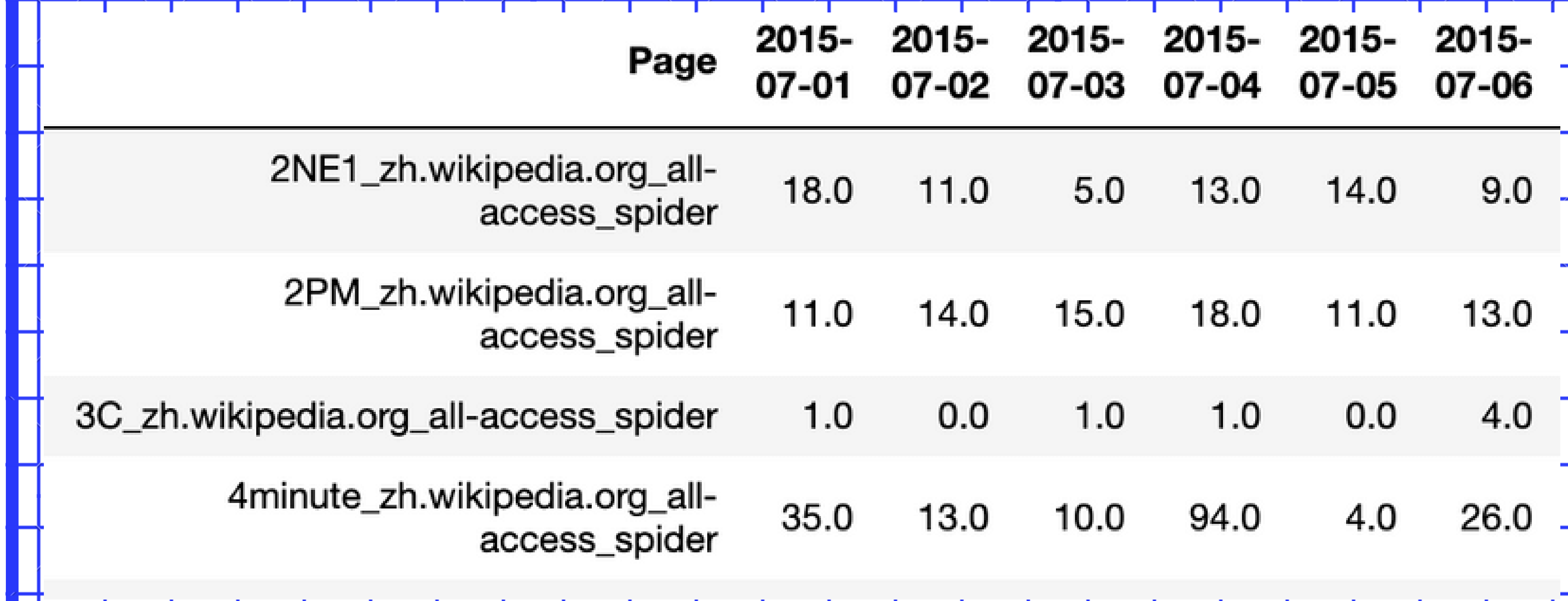
**Source of data**: Kaggle.

**Format**: CSV.

**Shape:** (145063, 551).

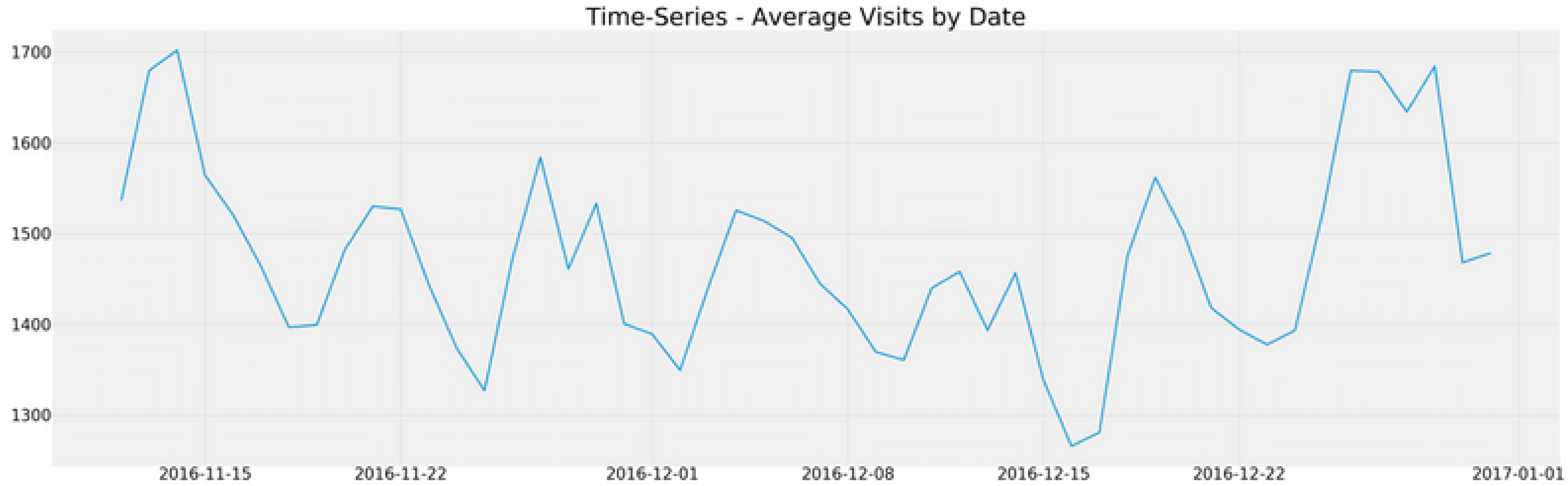
**Size:** 280.8 MB.

**Date:** 2015-2016-2017



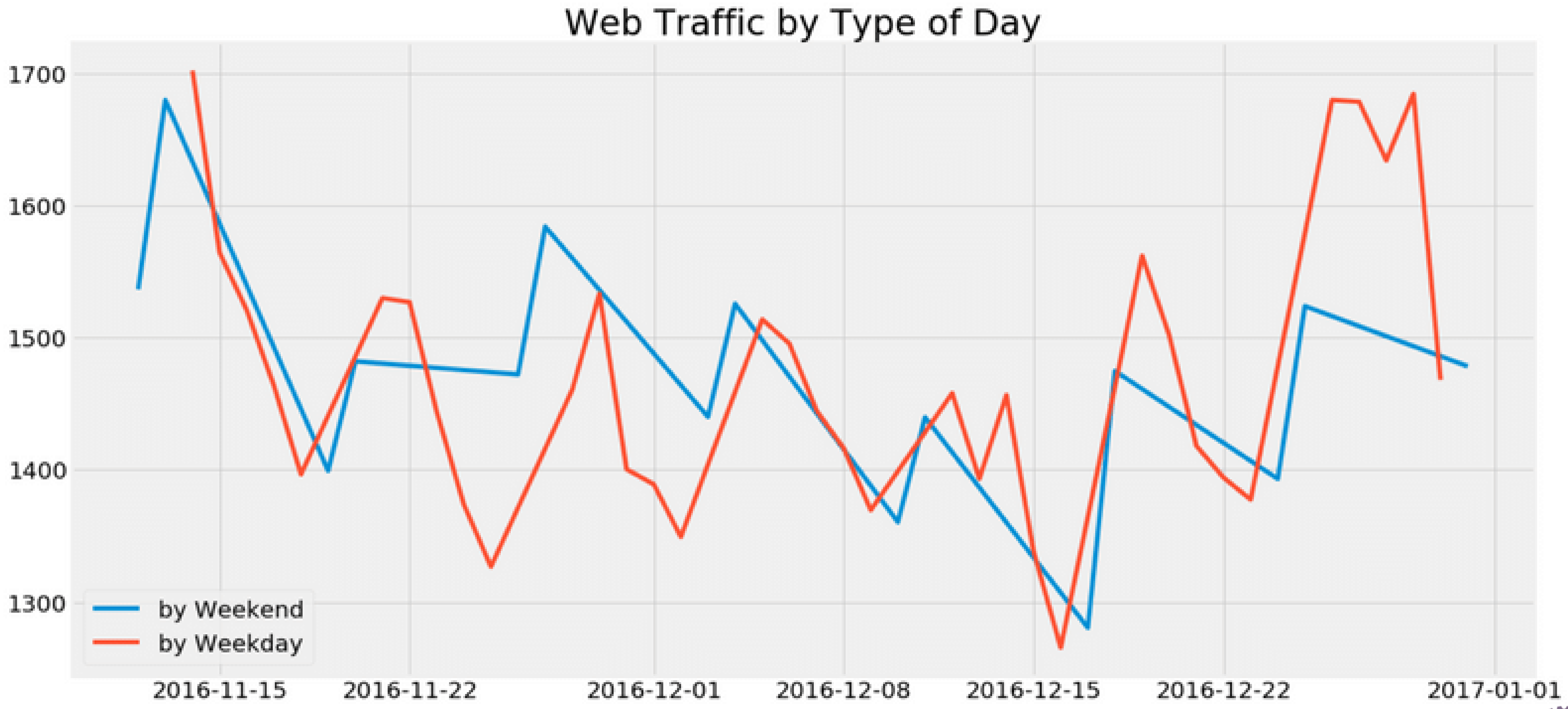
| Page | 2015-07-01 | 2015-07-02 | 2015-07-03 | 2015-07-04 | 2015-07-05 | 2015-07-06 |
|---|---|---|---|---|---|---|
| 2NE1_zh.wikipedia.org_all-access_spider | 18.0 | 11.0 | 5.0 | 13.0 | 14.0 | 9.0 |
| 2PM_zh.wikipedia.org_all-access_spider | 11.0 | 14.0 | 15.0 | 18.0 | 11.0 | 13.0 |
| 3C_zh.wikipedia.org_all-access_spider | 1.0 | 0.0 | 1.0 | 1.0 | 0.0 | 4.0 |
| 4minute_zh.wikipedia.org_all-access_spider | 35.0 | 13.0 | 10.0 | 94.0 | 4.0 | 26.0 |

# Basic EDA

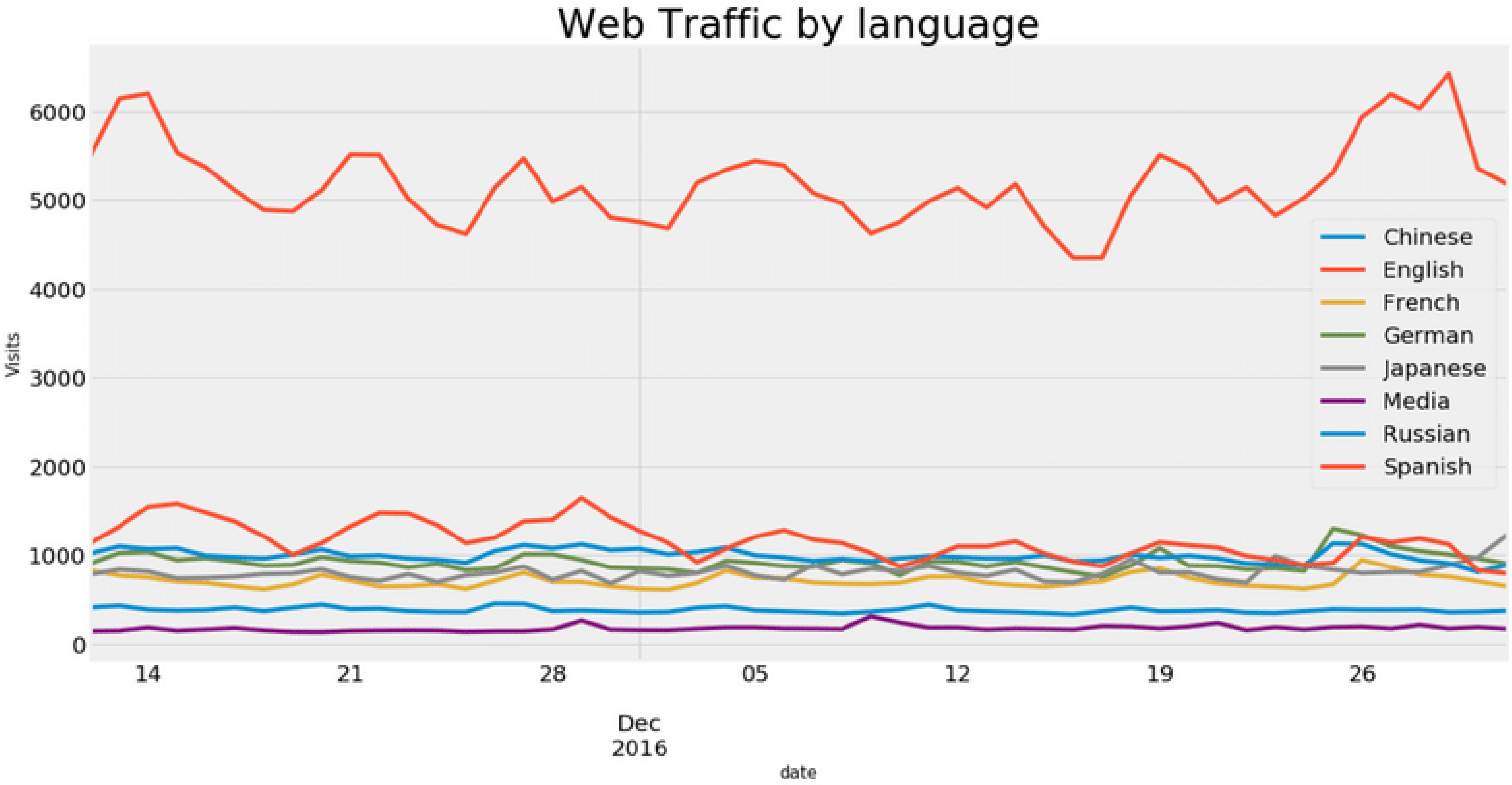


Time-Series - Average Visits by Date
1700
1600
1500
1400
1300
2016-11-15
2016-11-22
2016-12-01
2016-12-08
2016-12-15
2016-12-22
2017-01-01

# Basic EDA - Type of Day



# Basic EDA - Language



# Data Modeling

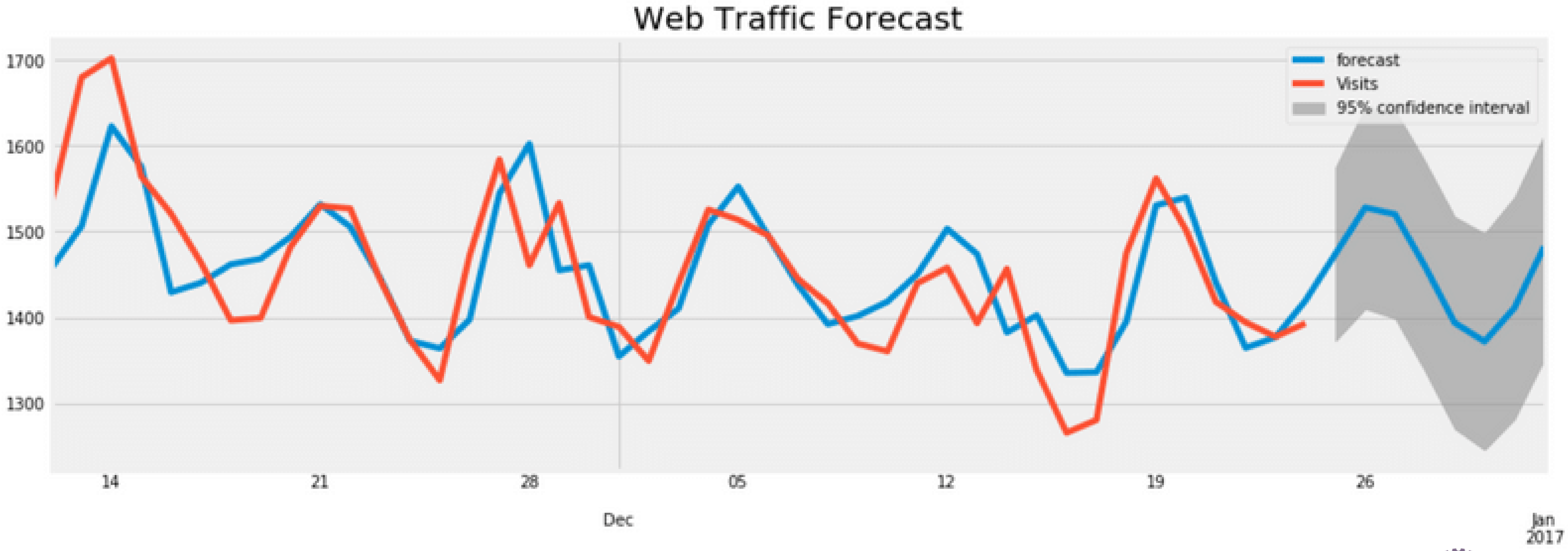
- Model used:

  ARMA

- Mean squared error(error rate):

  2662.5

# Predictions



# Challenges

- Missing data percentage was high: 42%

- Clean time-series data only covers two months

# FUTURE WORK

Explore the web traffic in terms of one language rather than all the languages