Data Mining Term Project 최종보고서

< 날씨와 배달데이터를 이용한 의사결정지원 모델>



**수업** : 데이터마이닝응용

**담당교수님** : 이석룡 교수님

**팀명** : 배달의 민조

**팀원** : 201703131 전종미

201701573 박지연

201502860 장동혁

201501703 송윤범

**<목차>**

한계점

1. 연구개요 및 목표
2. 데이터
   1. 데이터 수집방법
   2. 데이터 전처리
3. 배달음식 데이터 전처리
4. 날씨 데이터 전처리
5. 배달음식 데이터와 날씨 데이터 통합
6. 시간별 배달음식 데이터 전처리
7. 시계열 데이터 전처리
8. 사용할 최종 데이터
9. 사용할 DM 기능
10. 기초분석
11. 성별, 지점, 연령대, 요일과 총 배달건수의 관계
12. 날씨와 총 배달건수의 관계
13. 시계열 변수와 배달건수의 관계
14. 배달건수의 관계
15. 분석 결과
16. 의사결정나무
17. 신경망 분석
18. 시계열 분석
19. 기대효과
20. 느낀점

<한계점 및 어려웠던 점>

Raw 데이터가 개인별 데이터가 아니라 구별로 통합 되어있던 데이터였기 때문에 더 세세한 데이터를 알 수 없어 분석하는데 한계가 있었다. 또한 애초에 날씨데이터가 없는 날 자체가 존재하였고 attribute 중에서 null 값이 많은 attribute 들은 삭제했어야 했기 때문에 최종적으로 날씨 변수는 4가지밖에 사용하지 못하였다. 또한 우리가 생각했던 만큼 날씨 데이터가 배달건수에 많은 영향을 미치지 않은 것이 아쉬웠다. 또한 우리는 배달건수에 가장 많은 영향을 미치는 날씨 변수는 강수량이라고 생각했다. 강수량이 많으면 배달음식을 많이 시켜먹을 것이라고 예상했는데 1년치 데이터이다 보니 애초에 강수량이 ‘0’으로 없는 날이 대부분이였기 때문에 강수량 데이터 자체를 이분형으로 나눠야 했다. 따라서 비가 오냐 안오냐로 데이터를 전처리 할 수밖에 없었기 때문에 더욱 세세한 분석을 하지 못한 것이 한계점으로 생각한다. 신경망 분석에서 날씨변수를 구간화 하였을 때 그 구간의 기준이 애매하여 일반적으로 “춥다, 덥다, 바람이 세다, 비가 많이 내린다” 라는 식으로 설명하는데 어려움을 겪었다. 또한 시계열 분석에서 우리는 원래 데이터를 일별 및 요일별을 기준으로 계절성을 분석하고자 하였으나, n 값이 주별로 하면 52.18로, 일별로 하면 36.25로 정수가 아니였기 때문에 기준을 월별로 두고 파이썬으로 sarima 분석을 실시하였으나 error가 떴다. 따라서 우리 데이터에 알맞은 모수를 찾지 못했다고 생각하였다. R에는 auto arima라는 최적의 모수를 찾아주는 함수가 있었기 때문에 R로 arima분석을 실시하였다. 이러한 과정 중에서도 제대로 알지 못하고 우리가 가지고 있는 1년치 데이터에 맞춰 시계열 분석을 하였다. 여러가지 시계열 parameter(p, q, r, n)와 cut off 값과 lag 수, Window 설정값 등을 바꿔가면서 분석을 계속 실시했다. 여러가지 인터넷 검색과 시행착오 끝에 결국 1년치 데이터로 분석을 실시하였기 때문에 모델에서 error 발견되었다는 것을 깨달았다. Python과 R프로그램으로 시계열 분석을 더욱 자세히 분석하여 세세한 결과를 도출해 내고자 했지만 우리가 만족하는 결과에 미치지 않았기 때문에 아쉬웠다. 프로그래밍 하면서 코드 자체를 작성하는 데에도 어려움이 많았다. P-value, 신뢰구간과 같은 개념들은 통계 수업에서 학습하여 기초적인 분석에는 어려움이 없었지만 LAG과 window의 설정 기준 등 특정 기준을 우리 데이터에 맞춰 설정하는 것이 이번 term project를 하면서 가장 어려웠던 점이다. 앞으로 더욱 프로그래밍을 공부하여 알아갈 예정이다.

1. **연구개요 및 목표**

하루 평균 2500개의 자영업이 치열한 생존경쟁으로 인해 폐업이 되고 있는 현재, 이를 해결하기 위해 마케팅 컨설팅을 찾는 개인사업자들이 2011년(84만5235명)이후 5년만에 최고치를 기록하였다. 이에 따라 우리 팀은 우후죽순으로 생기는 배달음식점, 주문량의 부정확한 예측으로 인한 재료의 낭비, 날씨에 따른 주문량 차이를 고려하기 어려운 점 등과 같은 문제점이 경쟁시장에서 생존하기 위해서는 해결이 되어야 할 주요 문제점으로 인식을 하였다. 따라서 본 프로젝트는 여러가지 변수들과 배달 주문건수와의 관계를 파악해보고 날씨에 따른 배달음식 주문량의 변화를 예측하여 실제로 자영업자들이 활용 가능한 마케팅 전략을 제언하기 위해 본 모델을 구축해나갈 예정이다.

1. **데이터**
   1. 데이터 수집방법

**기상자료개방포털**에서 서울지역으로 한정하여 2018년 10월부터 2019년 9월에 해당하는 방재기상데이터를 통해 기온, 풍속, 기압 등과 같은 데이터를 수집하였다. 또한 SK telecom에서 제공하는 **BIG DATA HUB**에서 같은 기간 동안에 해당하는 배달음식 데이터를 수집하였는데, 배달 업종별 데이터에서 성별과 나이대가 나타나 있지 않은 업종도 존재하여, 이 attribute까지 나타내는 치킨, 중국집, 피자에 관련한 데이터만 따로 수집하였다. 또한 시간에 따른 총 배달건수를 각 배달음식에 관하여 나타내기 위해 구별로 나누어진 것이 아닌 시간이 명시되어 있는 서울시 전체 통합 배달 데이터를 따로 수집하였고 시계열 분석을 위하여 4년치 치킨, 피자, 중국음식 데이터를 수집하였다.

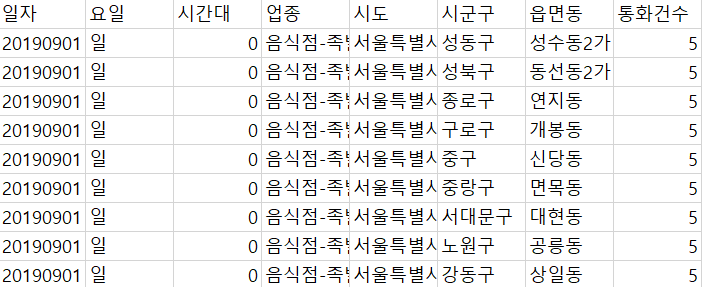
< **기상자료개방포털 –** 날씨데이터**>**



<**SK Telecom BIG DATA HUB –** 구별 배달건수 데이터>



<**SK Telecom BIG DATA HUB –** 시간별 배달건수 데이터>

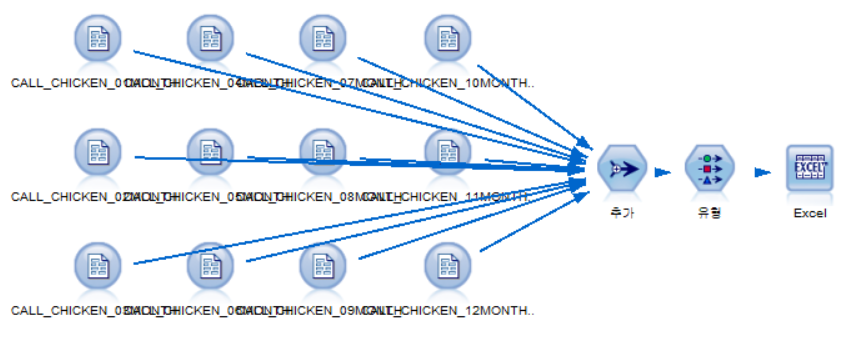


* 1. 데이터 전처리

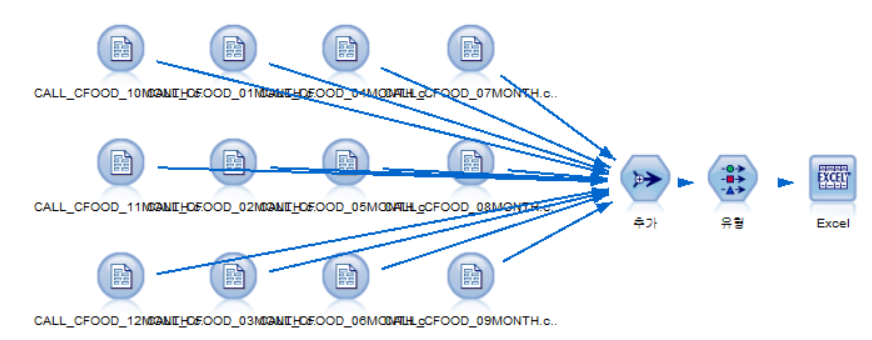
1. 배달음식 데이터 전처리

치킨, 중국음식, 피자 데이터가 각각 월별로 다른 파일로 수집되었기 때문에 추가 노드를 사용하여 하나의 파일로 통합하였다.

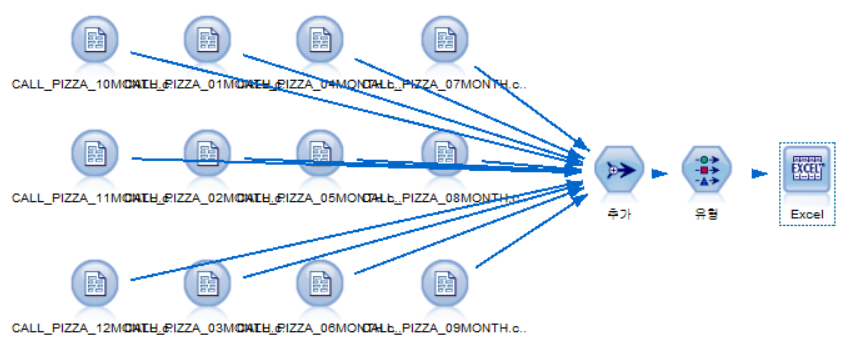
<월별 치킨 데이터 통합>



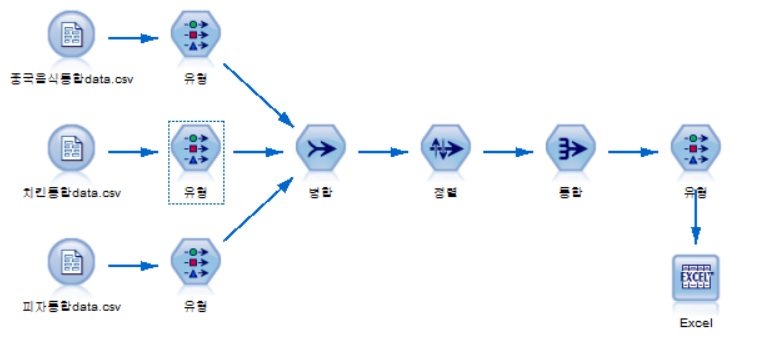
<월별 중국음식 데이터 통합>



<월별 피자 데이터 통합>

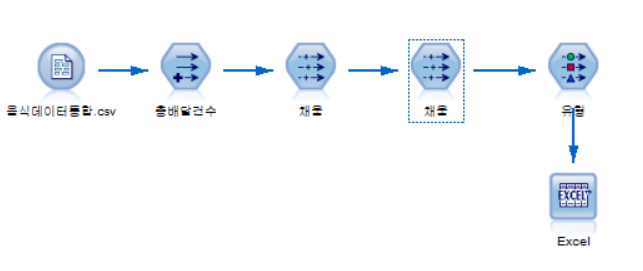


각 음식 데이터들의 통화건수 변수를 중국음식, 치킨, 피자통화건수로 변수명을 바꾸었고 이 데이터들을 통합하기 위해 병합노드를 사용하였다. 또한 읍면동으로 나누어져 있는 데이터를 구별로 나타내기 위하여 통합노드를 사용하여 전처리하였다.



중국음식, 치킨, 피자의 통화건수를 통합한 전체 배달음식 데이터와 날씨 데이터와의 관계를 파악하기 위해 파생노드를 사용하여 총 배달건수 attribute를 만들었고 데이터를 병합했을 때 생기는 null 값을 ‘0’ 으로 처리하기 위해 채움노드를 사용하였다.

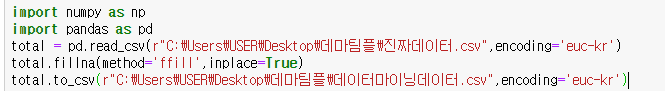
* 배달음식 raw 데이터를 봤을 때 배달건수가 0인 것은 존재하지 않고 데이터가 없는 경우 행 자체가 존재하지 않기 때문에 배달음식을 시키지 않아 데이터가 없는 것으로 판단하여 null 값을 ‘0’ 으로 설정하였다.



1. 날씨 데이터 전처리

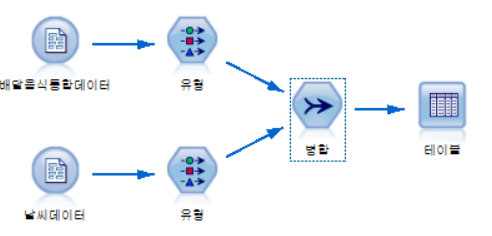
날씨 데이터와 배달음식 데이터의 통합을 위해 숫자로 나와있는 지점 code를 배달음식 데이터와 같이 구명으로 전환하였다. 날씨 데이터는 관측소 별로 나누어져 있어 기상청, 한강, 북악산 등 구에 관한 데이터가 아닌 것들은 분석 목표에 맞지 않기 때문에 사용할 데이터에서 제외하였다.

날씨 데이터의 null 값 처리를 위해 파이썬 코드를 사용하였다.



1. 배달음식 데이터와 날씨 데이터 통합

전체 데이터를 한번에 나타내기 위해 배달음식 데이터와 날씨 데이터를 통합하였다.

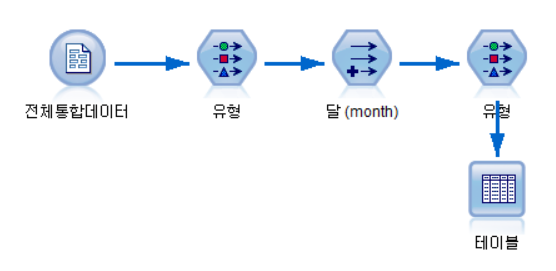


또한 분석에 사용하기 위해 추가적으로 일교차 변수, 시계열 변수인 달(month)를 생성하였고 사용하지 않을 변수를 제거하여 데이터를 생성하였다.

<일교차 변수 생성>

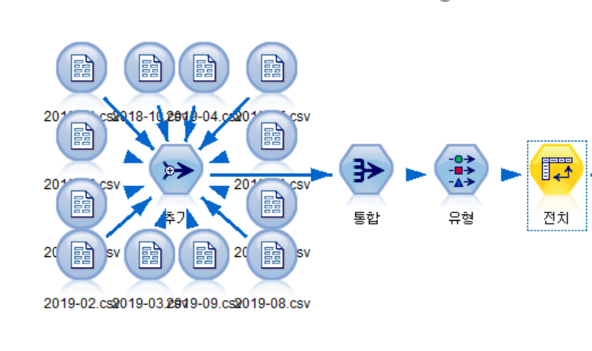


<달(month) 변수 생성>



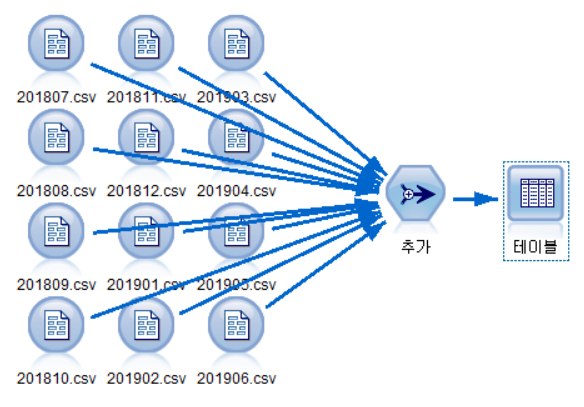
1. 시간별 배달음식 데이터 전처리

시간에 따른 배달음식 건수를 배달 음식별로 나타내기 위해서 월별로 데이터를 통합하였고 전치노드를 사용하여 업종 attribute로 되어있는 배달음식 데이터를 각 배달음식 데이터별로 attribute를 생성하였다.

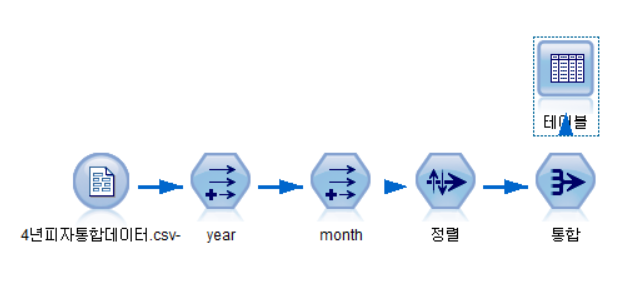


1. 시계열 데이터 전처리

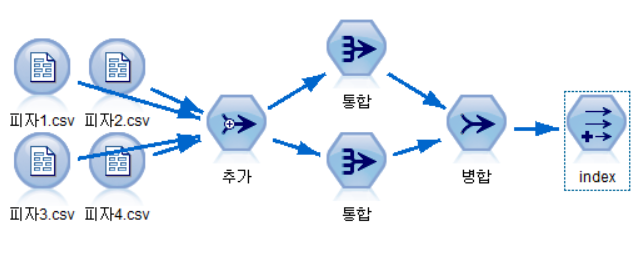
시계열 분석을 위한 치킨, 피자, 중국음식의 4년치 데이터를 통합하기 위해 추가노드를 사용하였다.



시계열 분석에 필요한 year, month 변수를 생성하였고 필요하지 않은 데이터들은 삭제하여 최종 시계열 데이터를 생성하였다.



시계열 분석에 필요한 index 값을 계산하기 위하여 파생노드를 사용하여 index 변수를 생성하였다.



* 치킨과 중국음식도 같은 방식으로 시계열 데이터를 생성하였다.

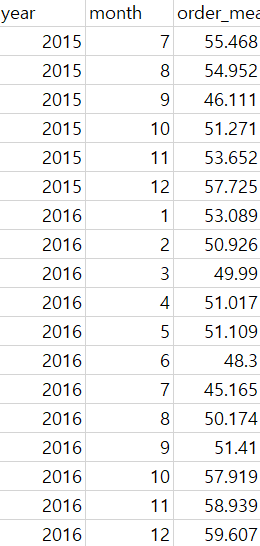
1. 사용할 최종 데이터

<의사결정 나무와 신경망 분석에 사용할 날씨 + 배달데이터>

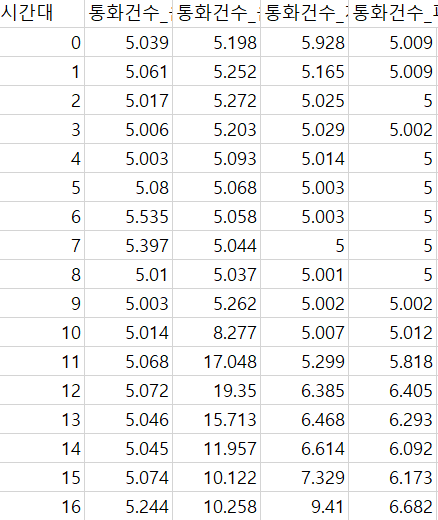


* 최종 데이터를 만든 후 분석에 사용하지 않을 최저 기온 시각(hhmi), 최고 기온 시각(hhmi), 최대 순간 풍속(m/s), 최대 순간 풍속 시간(hhmi), c최대 순간 풍속 풍향(deg) 변수를 삭제하였다.
* 최종적으로 각 행의 치킨, 피자, 중국음식 배달건수를 평균낸 총 배달건수의 평균, 일교차, 달(month) 변수가 추가 되었고 시간 순서대로 정렬하였다.

<시계열 데이터>



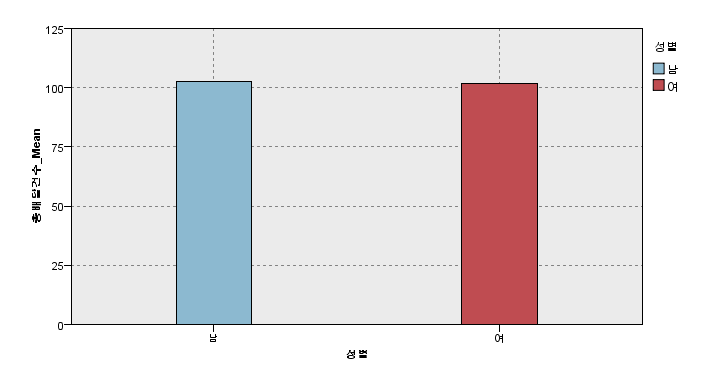
<시간대별 배달 건수 데이터>



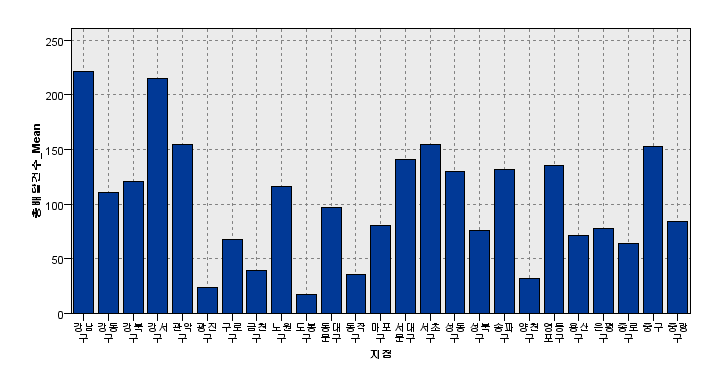
1. 사용할 DM 기능

|  |  |
| --- | --- |
| **사용할 DM 기능** | |
| **의사결정나무** | 배달 주문건수에 영향을 미치는 주요 attribute 선정 |
| **신경망 분석** | 날씨 데이터를 이용한 배달 건수 예측 |
| **시계열 분석** | 시계열 데이터로 미래 추이 예측 |

1. 기초분석
2. 성별, 지점, 연령대, 요일과 총 배달건수의 관계



* 성별에 따른 총 배달건수의 평균은 차이가 미비함을 보여준다.



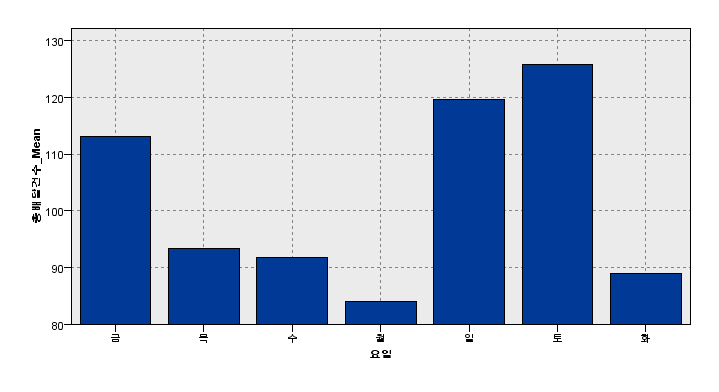
* 구에 따른 총 배달건수의 평균은 위와 같은 분포를 보여주고 있다.

강남구, 강서구에서는 높은 배달건수를 보이며

광진구, 도봉구에서는 낮은 배달건수를 보여주고 있다.

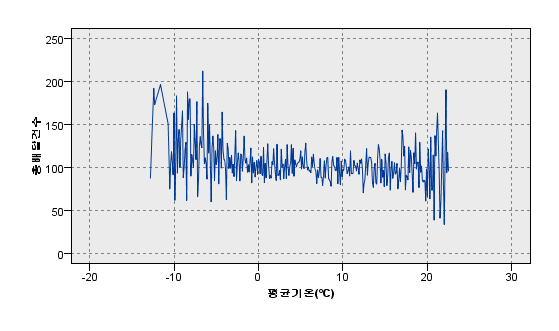


* 연령대에 따른 총 배달건수의 평균은 40대가 가장 높고 10대와 60대는 상대적으로 낮은 평균을 보여준다.

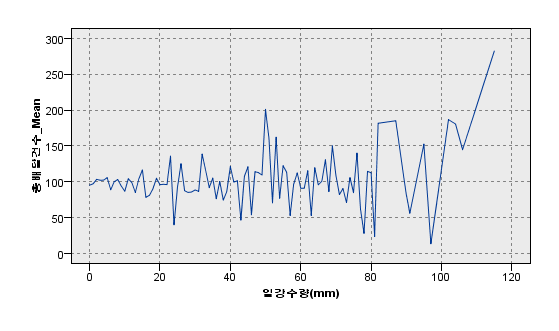


* 요일에 따른 총 배달건수의 평균은 금요일과 주말이 주중에 비해 특히 높은 것을 알 수 있다.

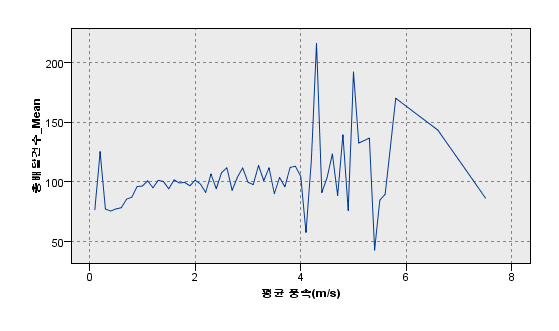
1. 날씨와 총 배달건수의 관계



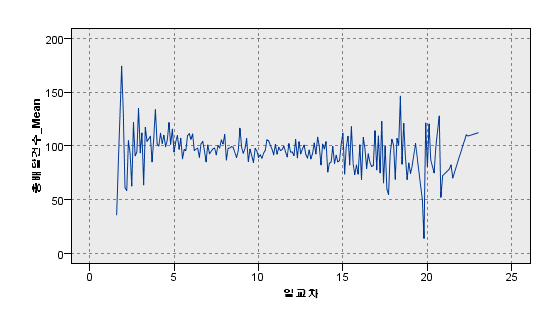
* 평균기온이 높거나 낮을 때 총 배달건수가 많아지는 추세를 보이고 그 외 기온에서는 특정 범위내의 수치를 유지하고 있다.



* 100mm 이상의 강수량에서 총 배달건수가 증가하는 추세를 보인다.

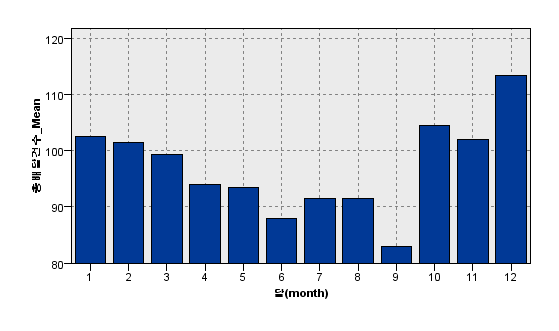


* 4 ~ 6 m/s 에서 총배달건수가 많은 풍속이 있지만 6 m/s 이상의 풍속에서는 감소하는 추세를 보인다.

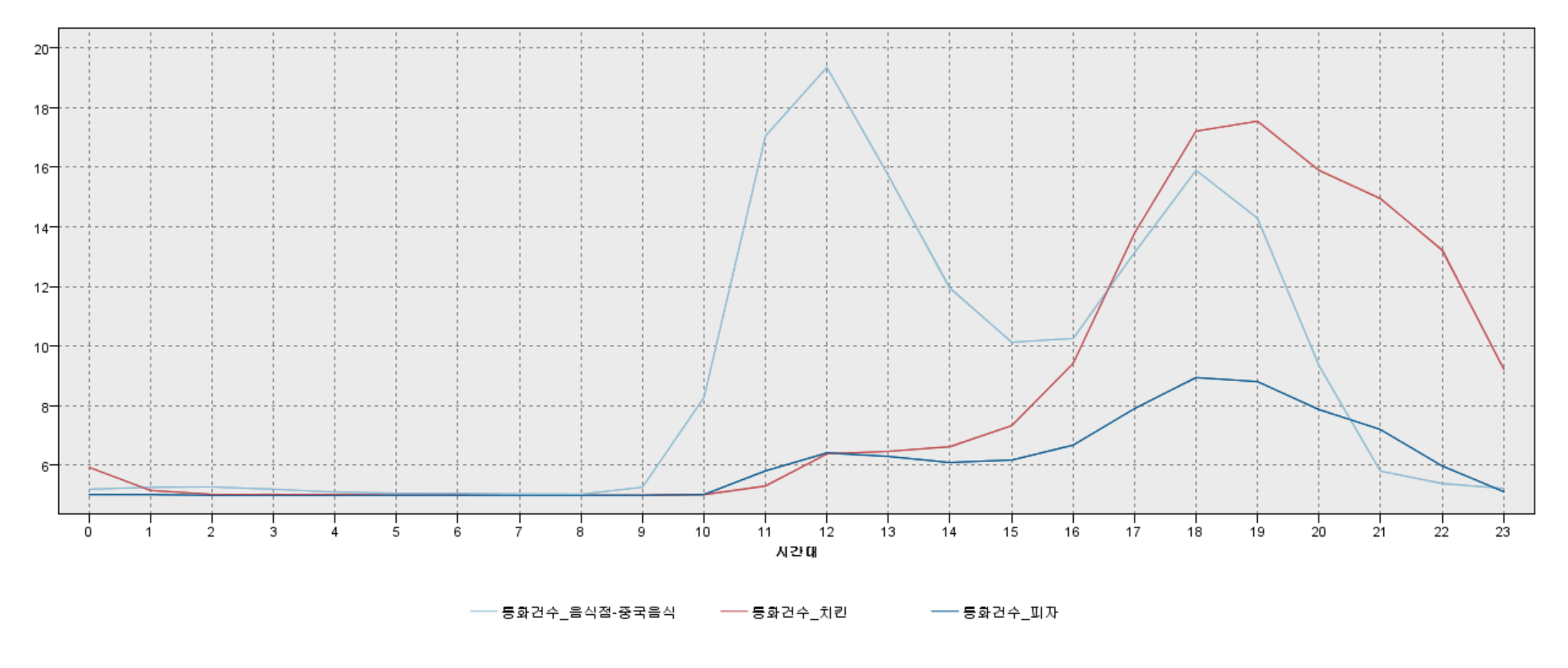


* 일교차가 커 저녁시간에 추워지게 되면 배달건수가 많아질 것이라는 예상과 달리 별 다른 상관관계를 보여주지 못하였다.

1. 시계열 변수와 배달건수의 관계



* 연말에 상대적으로 총 배달건수가 높은 것을 알 수 있다.

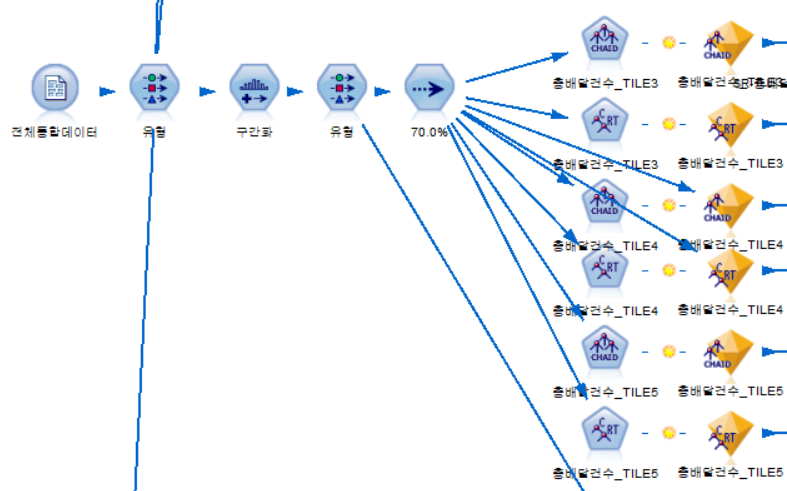


* 중국음식은 점심과 저녁시간 모두 배달건수가 많은 반면 치킨과 피자는 저녁시간에만 배달건수가 많은 것을 볼 수 있다. 피자는 두 음식과 달리 상대적으로 적은 판매를 보이고 있다.

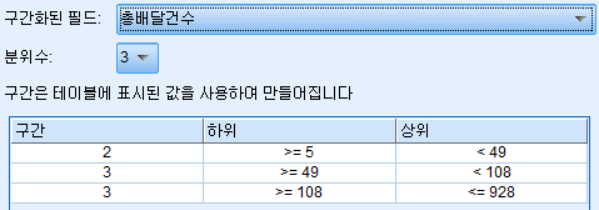
결과적으로 성별에 따라서는 뚜렷한 차이를 보이지 않고 강남구, 강서구의 경우 평균적으로 200건이 넘는 배달건수를 나타내고 있으나 광진구, 금천구, 도봉구, 동작구, 양천구와 같은 경우 현저히 낮은 수치를 보이고 있다. 따라서 광진구, 금천구, 도봉구, 동작구, 양천구의 배달 이용률을 증진시키는 방안을 모색해야 한다. 또한 10대는 경제적 능력이 부족하고 60대는 배달 음식 서비스 이용이 어려울 것으로 예상되기 때문에 10대와 60대의 배달 이용률이 낮은 것을 볼 수 있다. 따라서 배달 이용률이 높은 40대, 30대의 고객 유지에 힘쓰고 10대와 60대는 각 연령대에 맞는 마케팅 방안을 제시해야한다. 요일에 따라서는 금요일과 주말에 많은 분포를 보이고 평일은 상대적으로 적으므로 평일에 배달 서비스 이용률을 높여야 한다. 배달업체는 시간을 고려하여 각 시간대에 적합한 배달음식 서비스를 제공할 수 있어야 하며 날씨와 연관 지어 위와 같은 마케팅 전략을 모색해야 할 것이다.

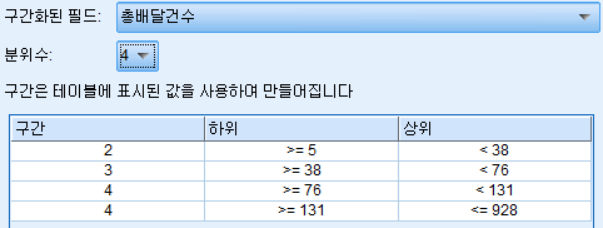
1. **분석 결과**
2. **의사결정나무**

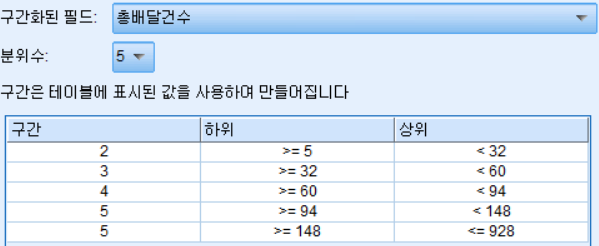
단순 그래프가 아닌 수치적 통계를 나타내 보고자 의사결정나무를 사용하여 어떤 변수가 총 배달건수에 가장 영향을 많이 미치는지를 분석하고자 한다.



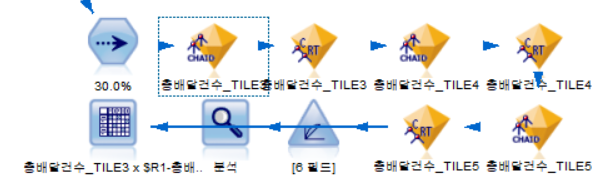
C&RT, CHAID 알고리즘을 사용하기 위해서 목표필드를 변환해야 한다. 구간화 노드를 통해 목표필드인 총 배달건수를 3분위, 4분위, 5분위로 구간화 하였다. 그리고 각각 지점, 성별, 요일, 연령대를 예측필드로 하여 Training Set (70%)에 대한 의사결정나무 분석을 실시하였다.

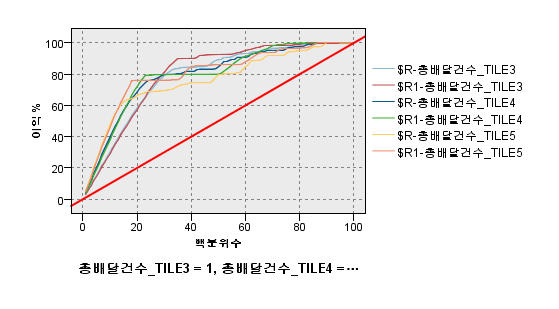




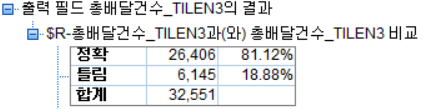


이후 Test Set(30%)에 모델을 적용하고 평가노드를 사용하여 예측 모델의 정확성을 비교하였다.





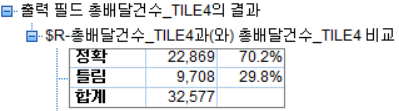
수치적으로 더 자세히 보여주기 위하여 분석노드를 사용하여 각 모델의 정확도를 비교해 보았다.



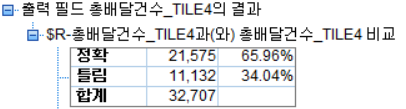
* 3분위 구간화 했을 때 C&RT\_ 총 배달건수의 정확도 : 81.12%



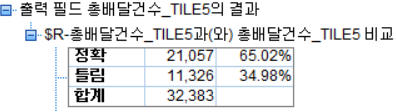
* 3분위 구간화 했을 때 CHAID\_ 총 배달건수의 정확도 : 77.06%



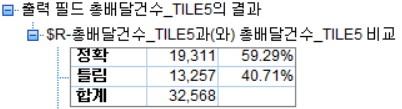
* 4분위 구간화 했을 때 C&RT\_ 총 배달건수의 정확도 : 70.02%



* 4분위 구간화 했을 때 CHAID\_ 총 배달건수의 정확도 : 65.96%

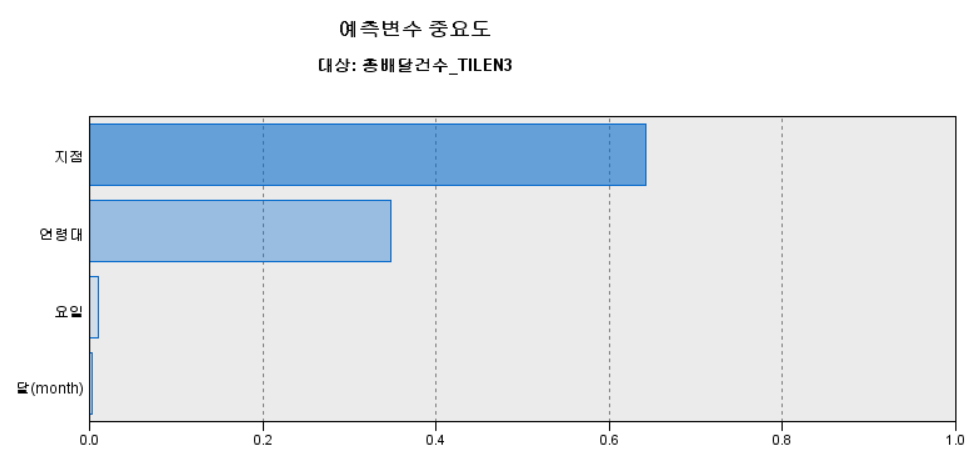


* 5분위 구간화 했을 때 C&RT\_ 총 배달건수의 정확도 : 65.02%



* 5분위 구간화 했을 때 CHAID\_ 총 배달건수의 정확도 : 59.29%

각각의 정확도를 비교했을 때 CHAID 알고리즘을 사용하는 것 보다 C&RT 알고리즘을 사용하는 것이 전체적으로 좋은 것을 볼 수 있다. 또한 총 배달건수를 3분위로 구간화한 것의 정확도가 81.12%로 가장 높으므로 3분위로 구간화한 C&RT 알고리즘을 사용하는 것이 적합하다. 이를 토대로 예측변수 중요도를 알아보고자 한다.

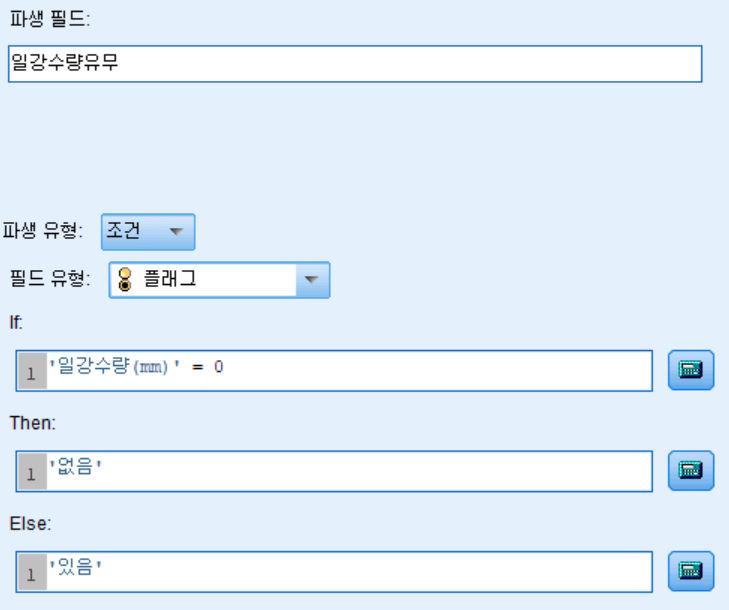


3분위로 구간화한 C&RT 알고리즘을 사용했을 때 예측변수 중요도는 지점과 연령대가 상대적으로 총 배달건수와 연관성이 높고 요일과 달(month)은 연관성이 낮음을 알 수 있다. 즉, 지점과 연령대를 고려하고 위에 제시한 도표그래프를 토대로 마케팅 방안을 제시해야 할 것이다.

1. **신경망 분석**

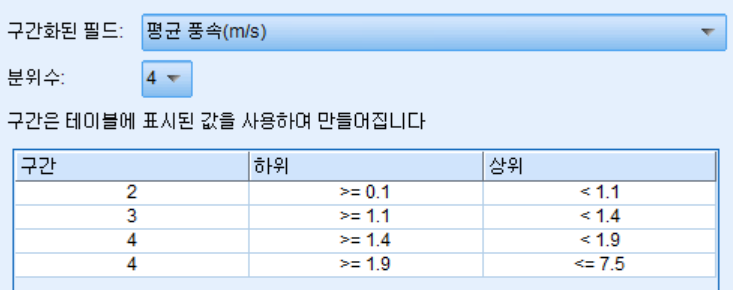
날씨데이터의 강수량, 풍속, 일교차, 평균기온을 예측필드로 하여 연속형 목표필드인 총 배달 건수를 각각 예측해보고자 한다. 먼저 예측필드 값들 중 강수량, 풍속, 일교차, 평균기온은 연속형이기 때문에 구간화를 사용하여 범주화를 시켰다.

<일강수량 구간화>

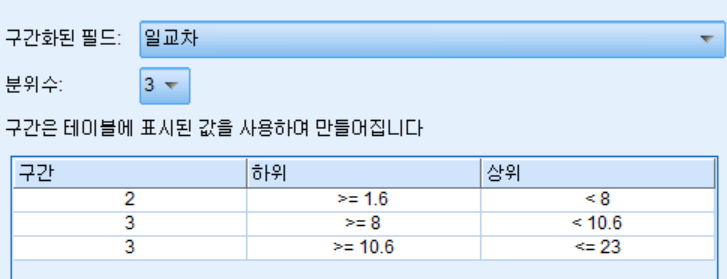


비가 내리지 않아 강수량 데이터가 ‘0’인 값이 많기 때문에 강수량은 구간화 하지 않고 이분형 데이터로 바꿔주었다.

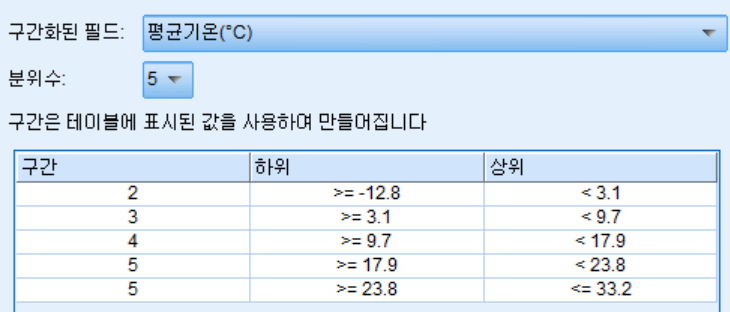
<풍속 구간화>



<일교차 구간화>



<평균기온 구간화>



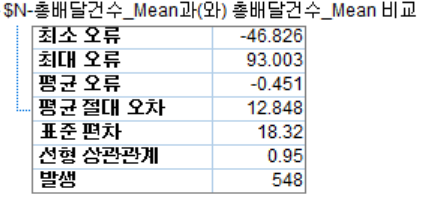
그 다음, Training Set (70%)에 대하여 신경망 분석 기법 MLP와 RBF를 이용해 모델을 생성하였다.



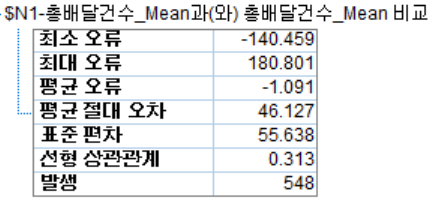
이후 Testing Set (30%)을 이용하여 모델의 적합성을 검증하였다.



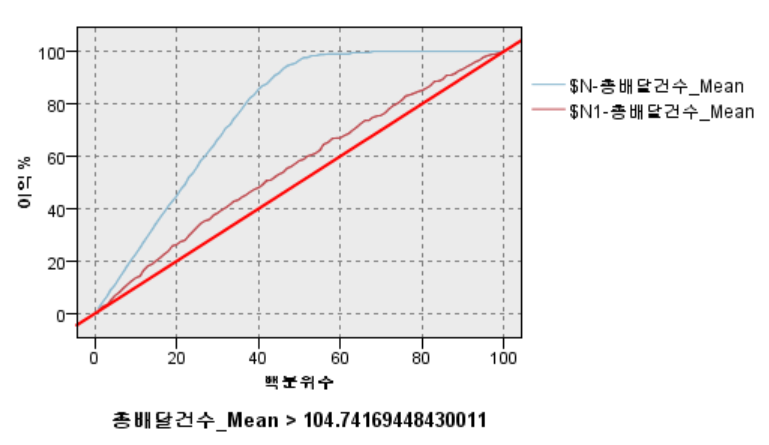
<MLP 기법>



<RBF 기법>



분석노드를 이용해 비교했을 때, 최대 오류 측면에서 MLP 기법의 모델이 수치가 더 적고, 선형 상관관계 측면에서는 MLP 기법의 수치가 0.95으로 RBF 기법의 0.3913에 비해 훨씬 1에 가깝다. 따라서 MLP 기법을 이용한 신경망 모델의 적합성이 더욱 뛰어나다.



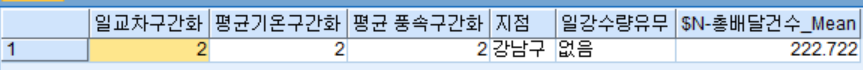
평가그래프를 이용해 비교해보아도 MLP 모델(파란색)이 기존 선형과 더 멀리 떨어져 있는 것을 볼 수 있다. 즉, MLP 기법이 RBF 기법보다 좋다고 말할 수 있다.

생성된 MLP 신경망 모델과 사용자입력 노드를 이용하여 새로운 날씨데이터의 강수량, 풍속, 일교차, 평균기온과 지점을 임의로 지정하여 총 배달건수를 예측해보도록 한다.

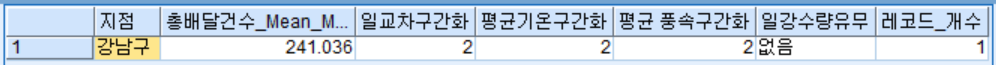




실제 값과 예측 값이 차이가 없는지 확인해보기 위해 위와 같이 임의로 변수를 넣었다.



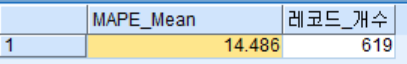
위의 사용자 입력 데이터에 따른 치킨 배달건수 평균의 예측값은 222.722이다.



데이터에서 해당되는 레코드만을 선택했을 때 실제 총 배달건수의 평균은 241.036으로 나왔다.

위의 예측값은 임의적으로 넣은 값이기 때문에 정확도가 떨어질 수 있다. 따라서 수치적으로 더 자세한 정확도를 알아보기 위해 MAPE를 사용하여 정확도를 검증하였다.

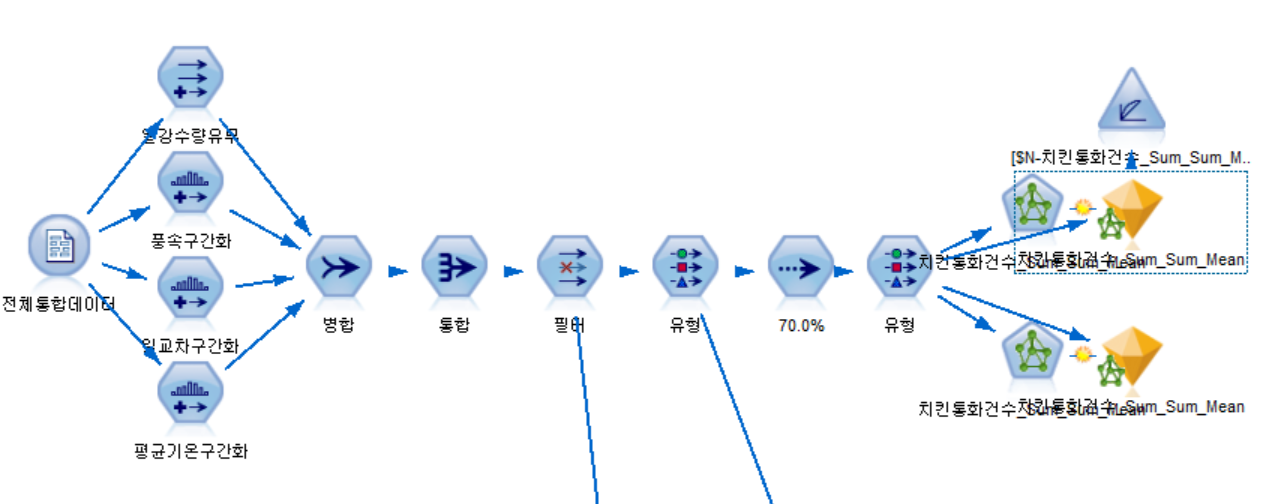
MAPE = (abs(실제 총 배달건수의 평균 – 예측 총 배달건수의 평균) / 실제 총 배달건수의 평균) \* 100



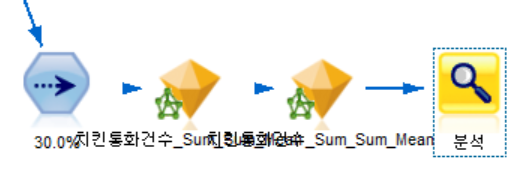
MAPE를 계산한 결과 14.486로 비교적 정확한 예측을 했다고 할 수 있다.

날씨와 총 배달건수와의 관계만이 아닌 자영업자한테 영업에 도움이 되는 더 세부적인 예측값을 제언해주기 위해 연령대와 성별을 추가적으로 예측변수에 넣고 예측 목표값을 치킨, 피자, 중국집으로 나눠 위와 같은 방법으로 더욱 세부적인 분석을 해 보았다.

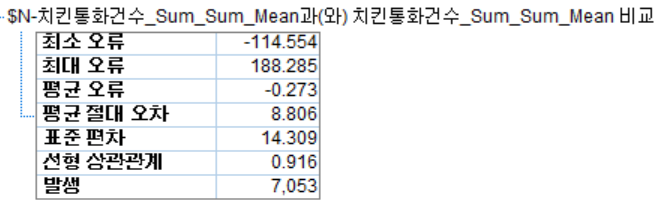
그 다음, Training Set (70%)에 대하여 신경망 분석 기법 MLP와 RBF를 이용해 모델을 생성하였다.



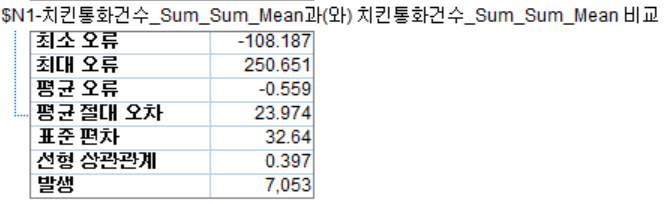
이후 Testing Set (30%)을 이용하여 모델의 적합성을 검증하였다.



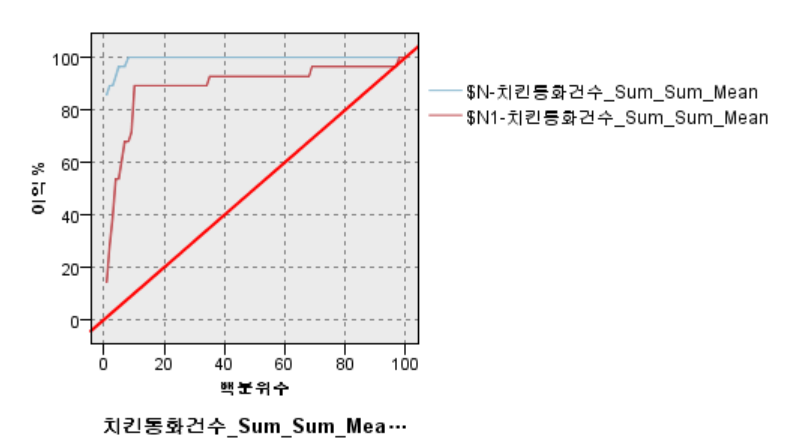
<MLP 기법>



<RBF 기법>



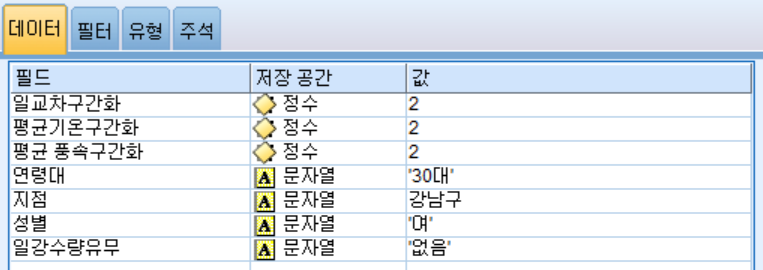
분석노드를 이용해 비교했을 때, 최대 오류 측면에서 MLP 기법의 모델이 수치가 더 적고, 선형 상관관계 측면에서는 MLP 기법의 수치가 0.916으로 RBF 기법의 0.397에 비해 훨씬 1에 가깝다. 따라서 MLP 기법을 이용한 신경망 모델의 적합성이 더욱 뛰어나다.



평가그래프를 이용해 비교해보아도 MLP 모델(파란색)이 기존 선형과 더 멀리 떨어져 있는 것을 볼 수 있다. 즉, MLP 기법이 RBF 기법보다 좋다고 말할 수 있다.

생성된 MLP 신경망 모델과 사용자입력 노드를 이용하여 새로운 날씨데이터의 강수량, 풍속, 일교차, 평균기온과 배달음식 데이터의 지점, 연령대, 성별을 임의로 지정하여 치킨 배달건수를 예측해보도록 한다.

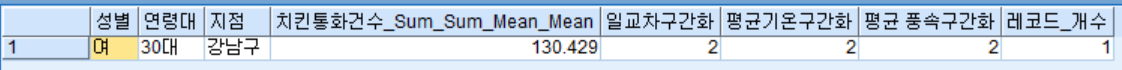




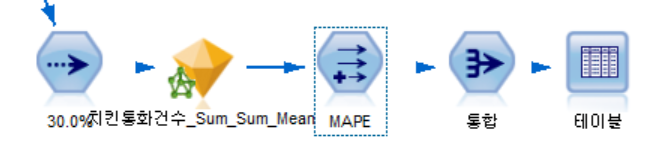
실제 값과 예측 값이 차이가 없는지 확인해보기 위해 위와 같이 임의로 변수를 넣었다.



위의 사용자 입력 데이터에 따른 치킨 배달건수 평균의 예측값은 114.204이다.



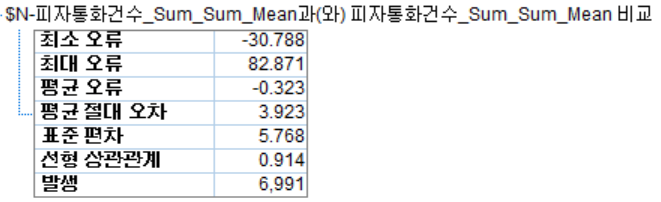
데이터에서 해당되는 레코드만을 선택했을 때 실제 치킨 배달건수의 평균은 130.429으로 나왔다.

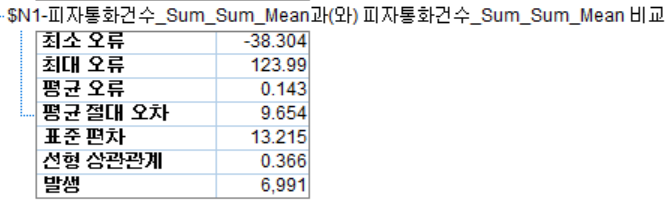


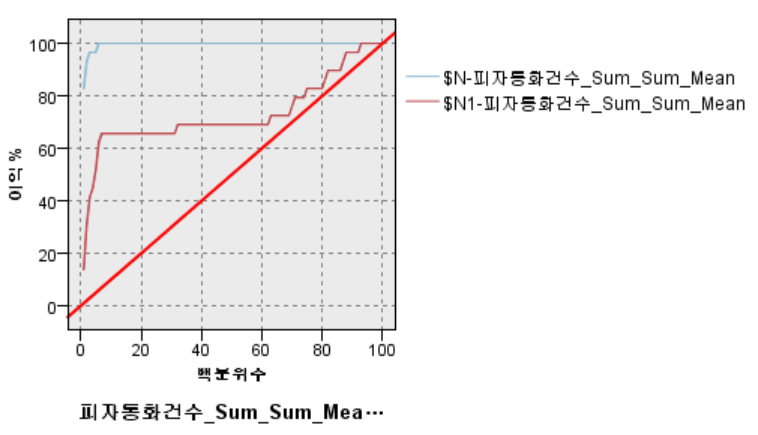


MAPE를 계산한 결과 30.541로 비교적 합리적 예측을 했다고 할 수 있다.

위와 같은 방식으로 피자와 중국음식 배달건수에 대한 신경망 분석을 하였을 때 MLP 기법이 적합하여 치킨과 똑같은 조건으로 피자와 중국음식의 배달 건수를 예측해 보았다.





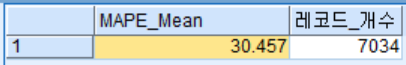




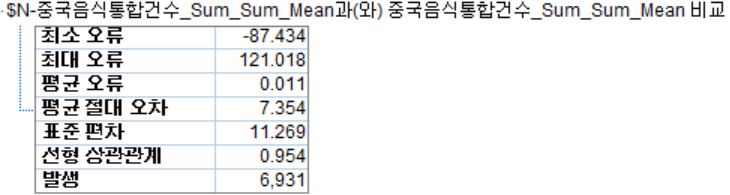
위의 사용자 입력 데이터에 따른 피자 배달건수 평균의 예측값은 46.857이다.

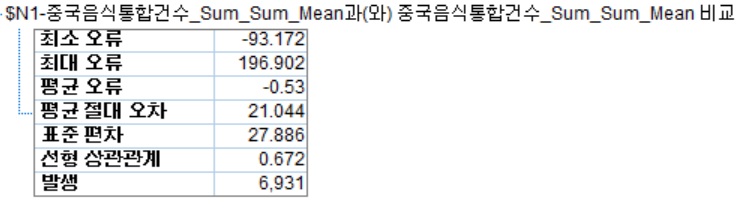


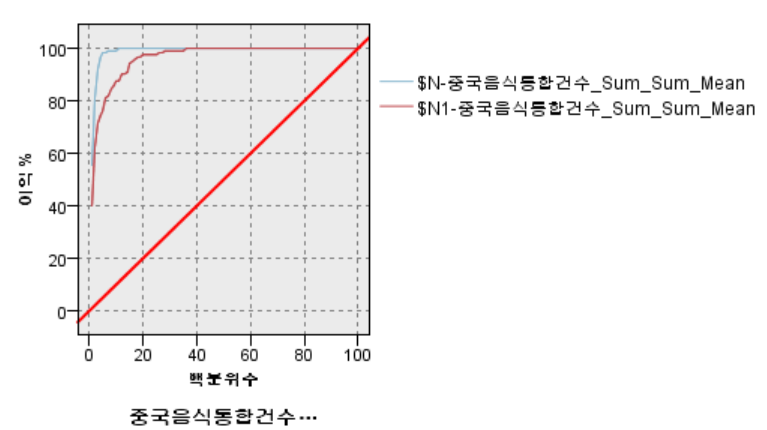
데이터에서 해당되는 레코드만을 선택했을 때 실제 피자 배달건수의 평균은 37.787으로 나왔다.



MAPE를 계산한 결과 30.457로 비교적 합리적 예측을 했다고 할 수 있다.

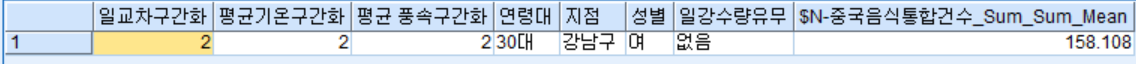








위의 사용자 입력 데이터에 따른 중국음식 배달건수 평균의 예측값은 154.571이다.



데이터에서 해당되는 레코드만을 선택했을 때 실제 중국음식 배달건수의 평균은 158.108으로 나왔다.



MAPE를 계산한 결과 30.301로 비교적 합리적 예측을 했다고 할 수 있다.

따라서 치킨, 피자, 중국음식의 배달건수를 예측하였을 때 모두 MAPE 값이 30 내외로 비교적 합리적 예측을 했다고 할 수 있다. 즉, 서울시 내의 여러 지점의 자영업자들은 이 예측 방식을 이용하였을 때 가상의 날씨에 따른 고객의 치킨, 피자, 중국음식 배달건수를 알 수 있고 연령대, 성별에 따라 알맞은 마케팅을 할 수 있을 것이다. 신경망 분석은 과과거의 모든 데이터를 분석할 필요가 없이 배달건수를 간편하게 예측해 볼 수 있다.

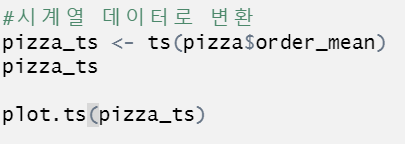
1. **시계열 분석**

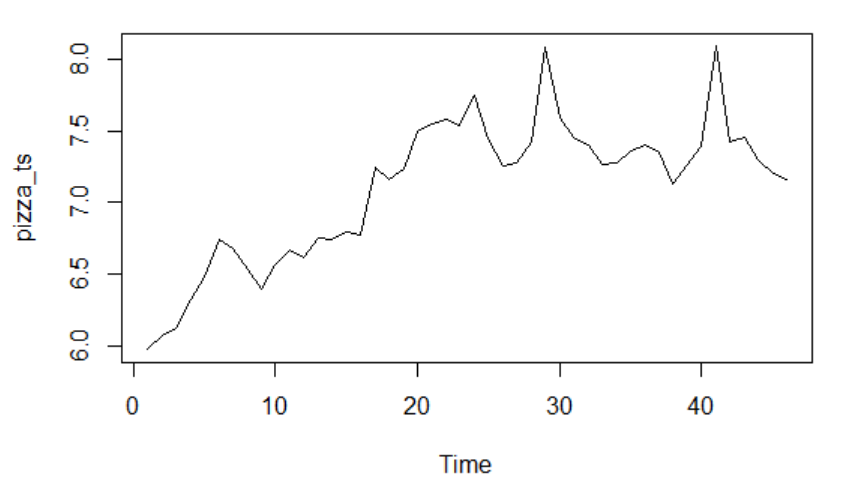
4년동안의 치킨, 피자, 중국음식의 배달건수 추세를 알아보고 향후 5달의 배달건수를 예측하기 위하여 R프로그램을 사용하여 시계열 분석을 하였다.

**<피자 배달건수 예측>**

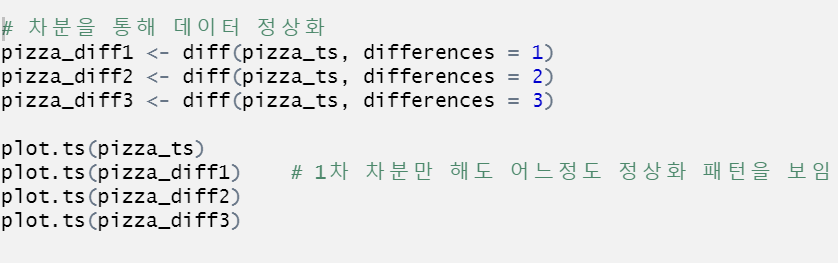


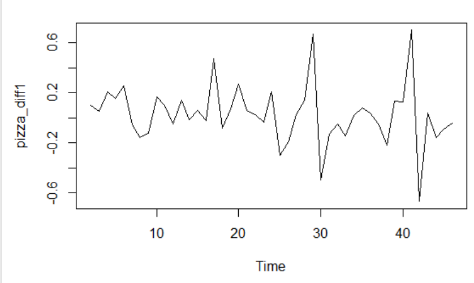
* 시계열 데이터로 변환



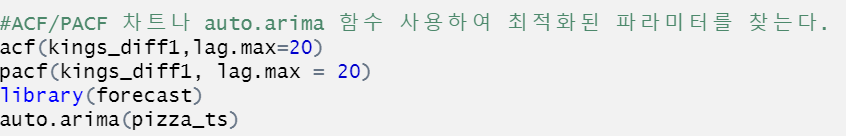


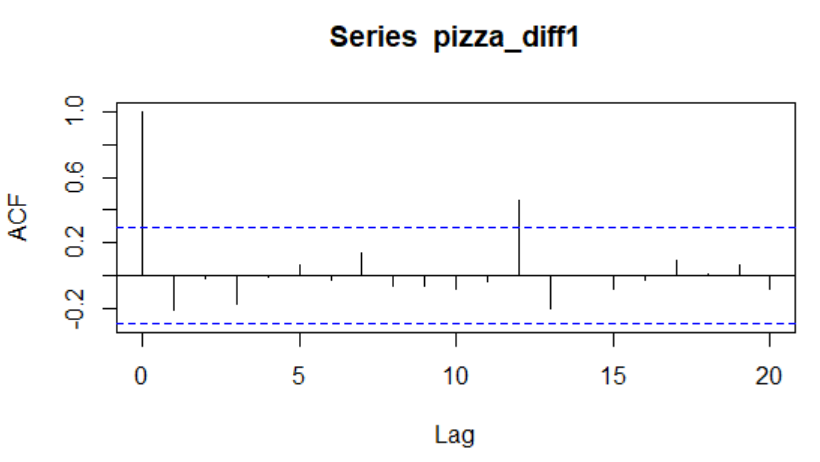
* 차분을 통해 데이터 정상화

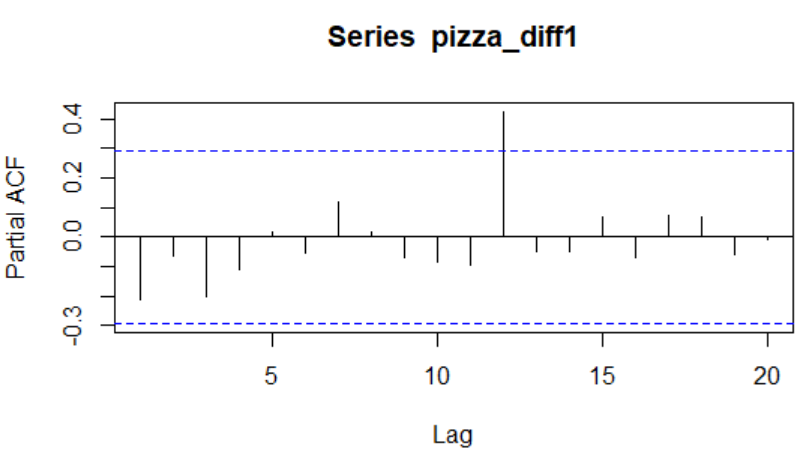


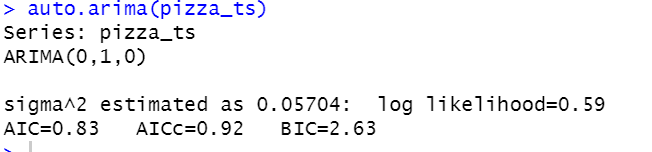


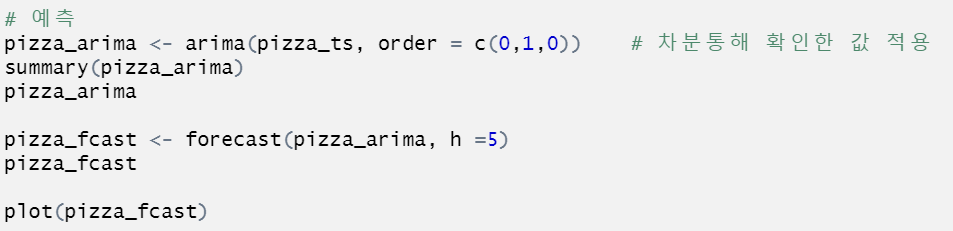
* ACF/PACF 차트와 auto.arima 함수를 사용하여 최적화된 파라미터를 찾음.

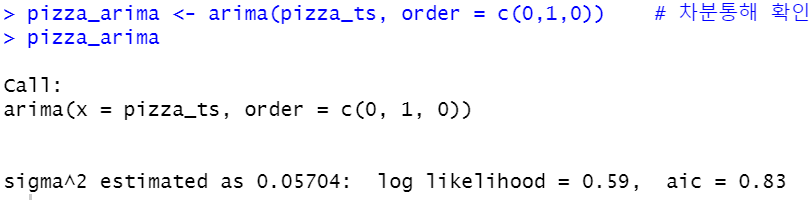


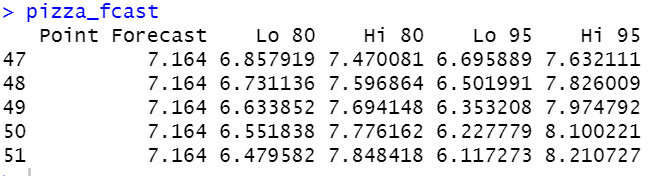




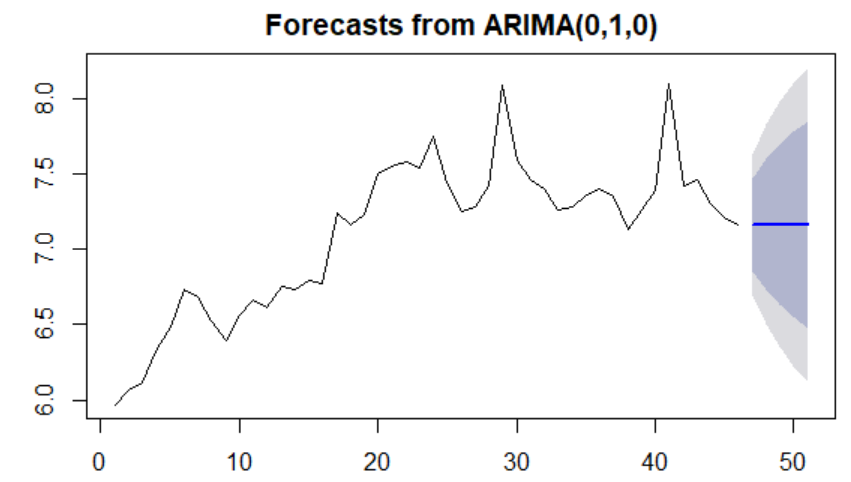




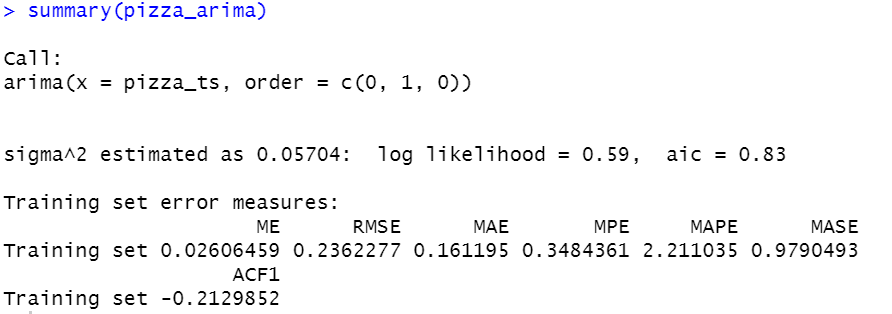




* 예측된 피자 배달건수 그래프



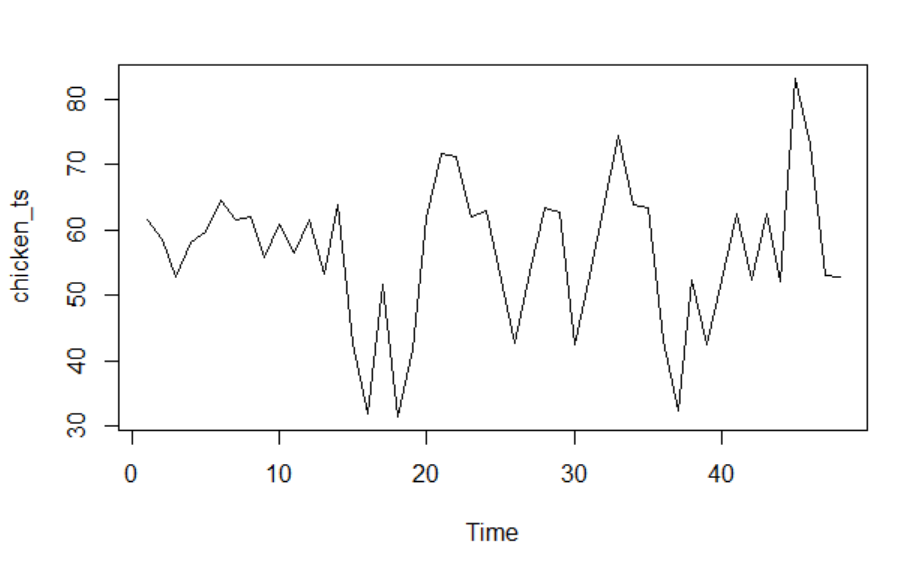
* 모델 평가



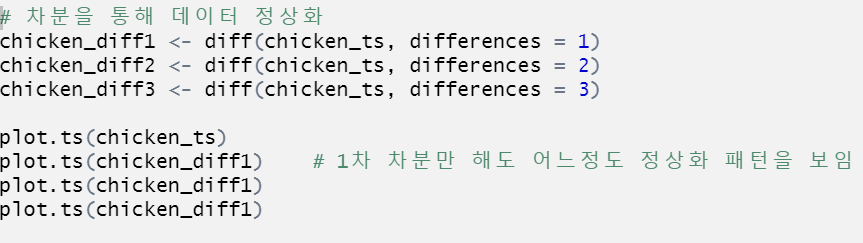
**->** MAPE를 봤을 때 2.21%로써 모델의 예측은 매우 정확하다고 볼 수 있다.

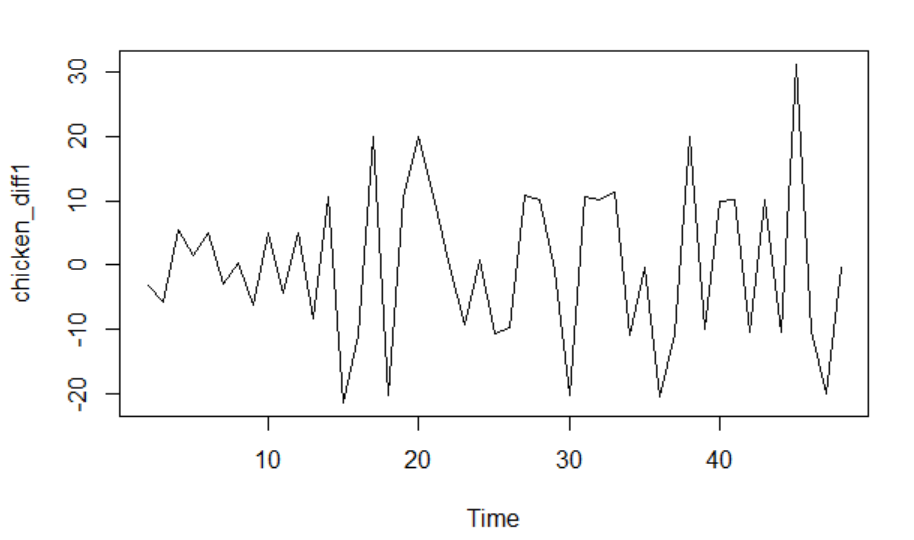
**<치킨 배달건수 예측>**

치킨도 위와 같이 데이터를 불러왔고, 시계열 변수로 변환하였다.

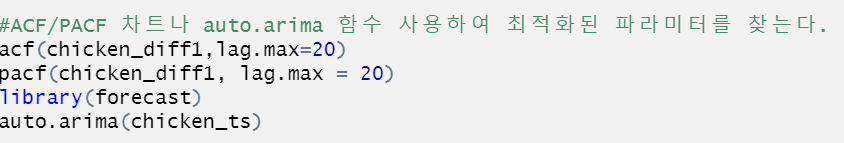


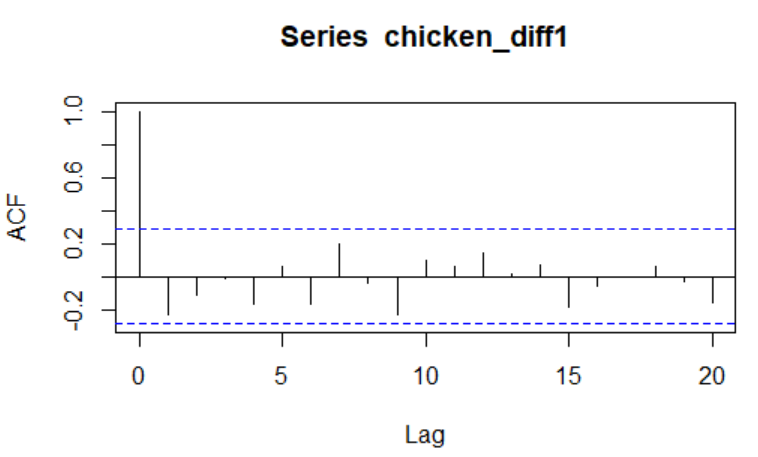
* 차분을 통해 데이터 정상화



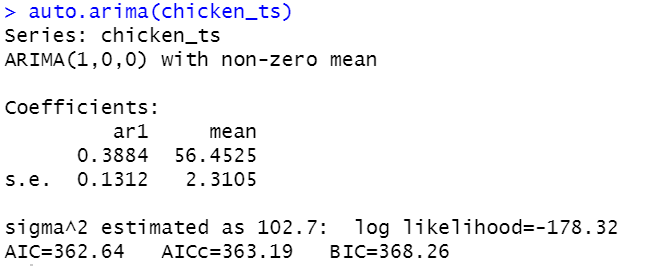


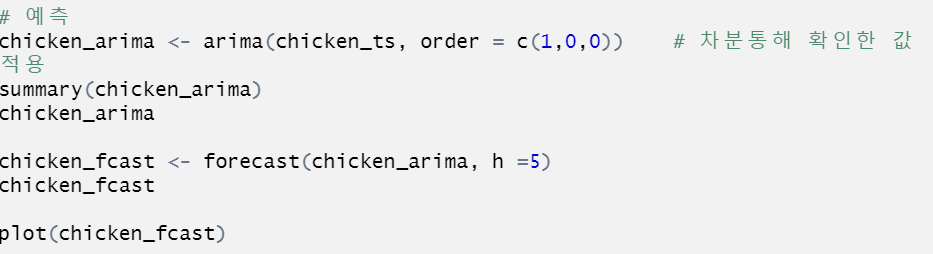
* ACF/PACF 차트와 auto.arima 함수를 사용하여 최적화된 파라미터를 찾음.

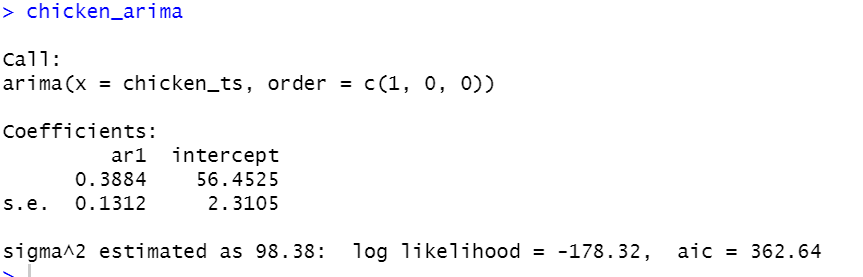






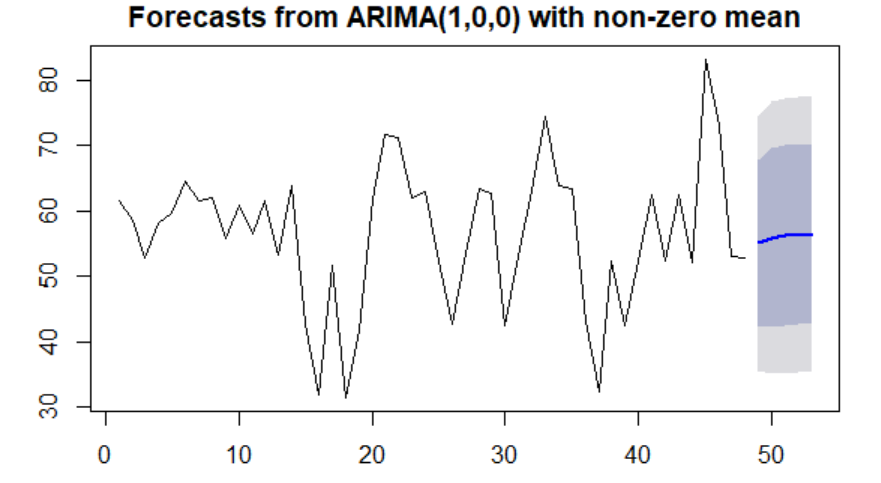




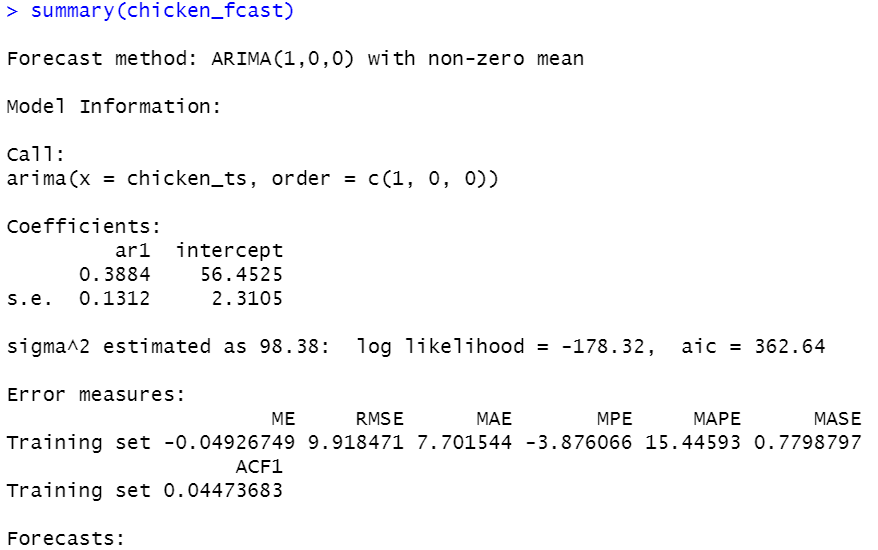




* 예측된 중국음식 배달건수 그래프



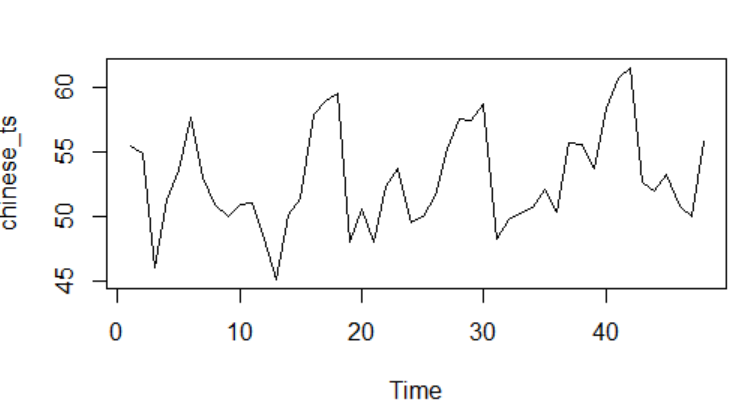
* 모델 평가



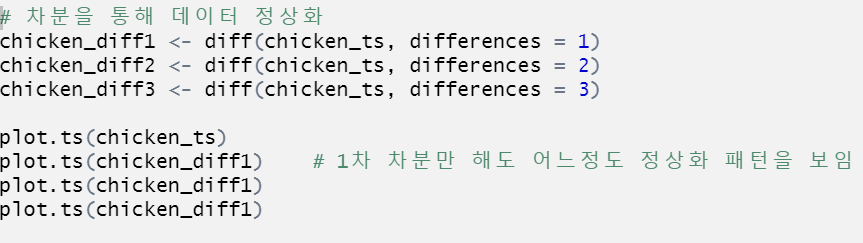
**->** MAPE를 봤을 때 15.4%로써 모델의 예측은 피자보다는 아니지만 비교적 정확한 예측이라고 볼 수 있다.

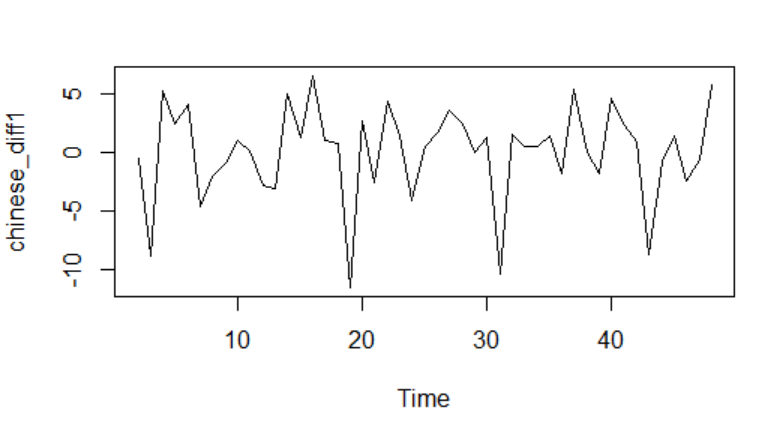
**<중국음식 배달건수 예측>**

중국음식데이터 또한 위와 같이 데이터 불러오고 시계열 변수로 변환했다.

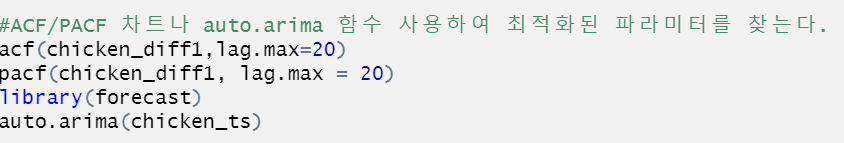


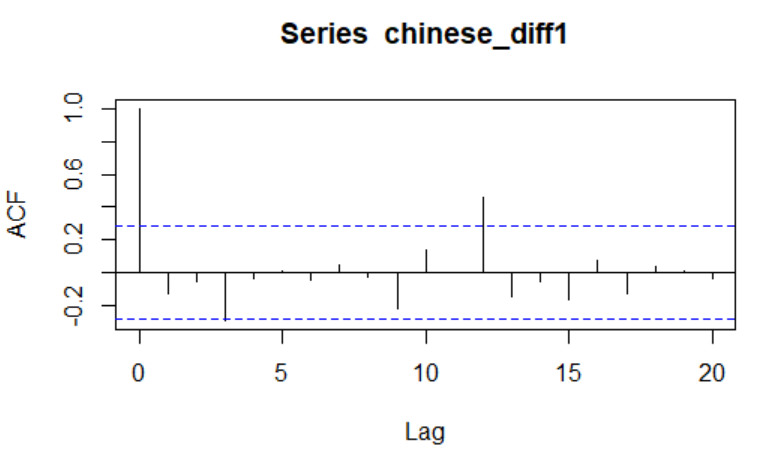
* 차분을 통해 데이터 정상화

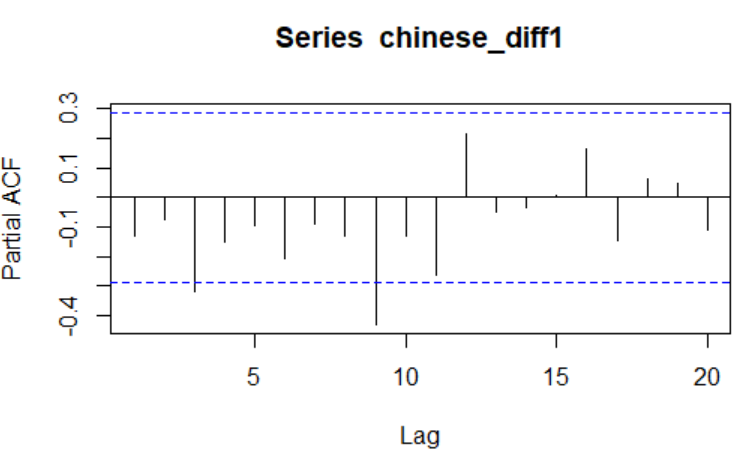


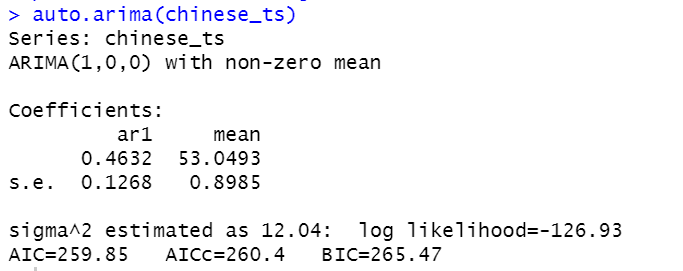


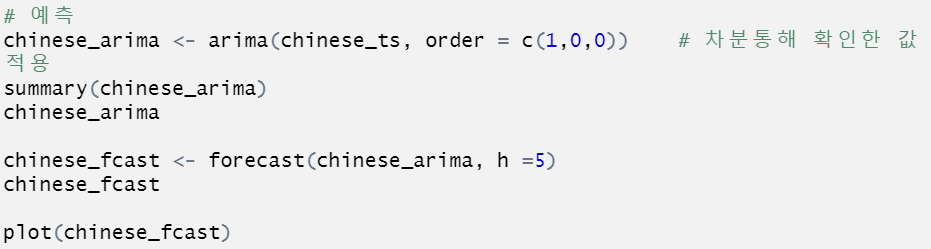
* ACF/PACF 차트와 auto.arima 함수를 사용하여 최적화된 파라미터를 찾음.

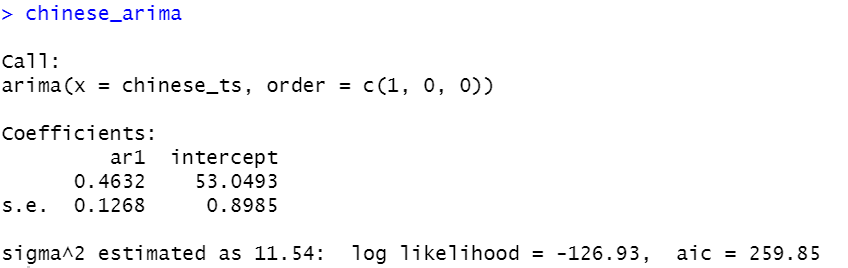


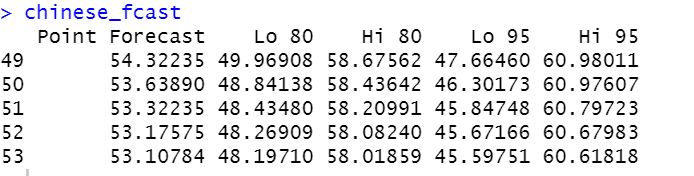




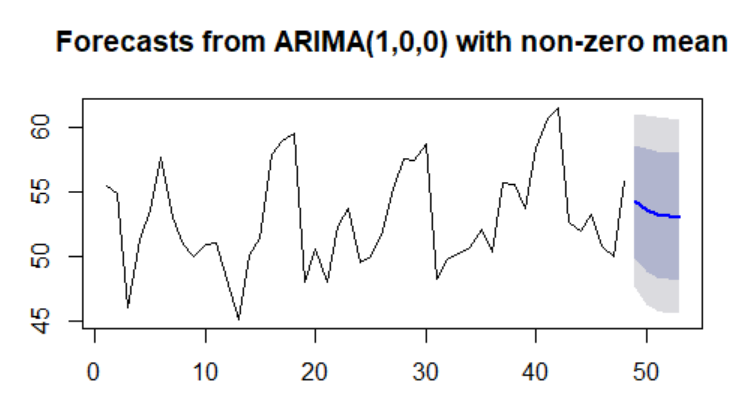




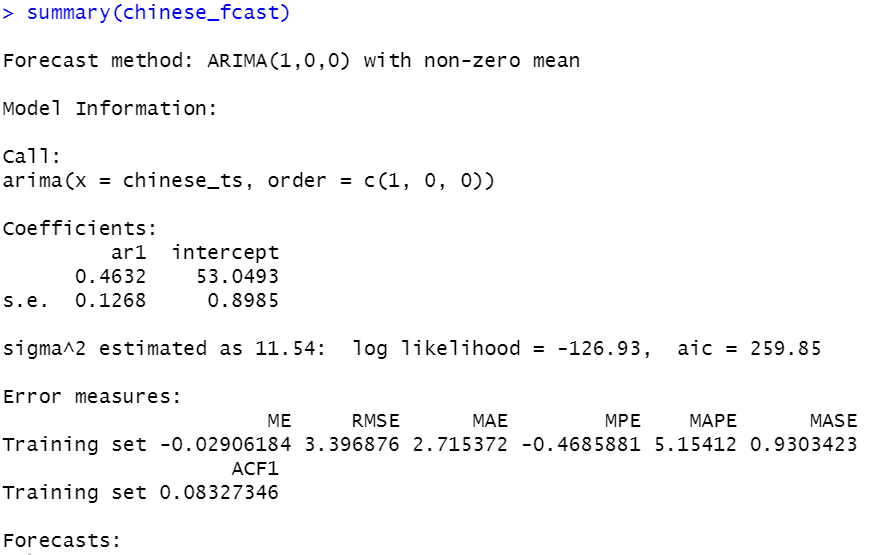




* 예측된 중국음식 배달건수 그래프



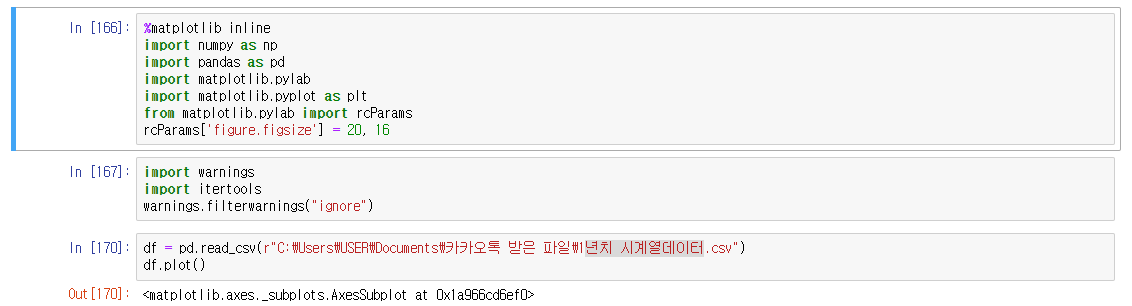
* 모델 평가



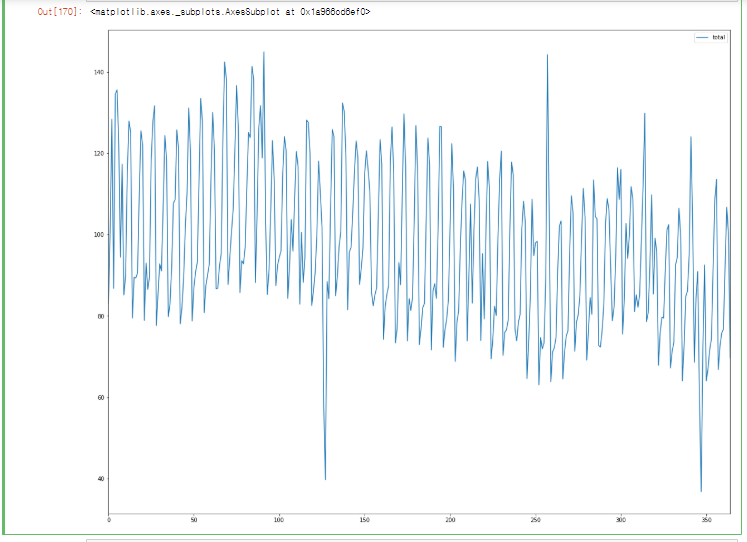
**->** MAPE를 봤을 때 5.15%로써 모델의 예측은 피자와 마찬가지로 10% 미만이기 때문에 매우 정확한 예측이라고 할 수 있다.

**+ SARIMA 분석 (With Python)**

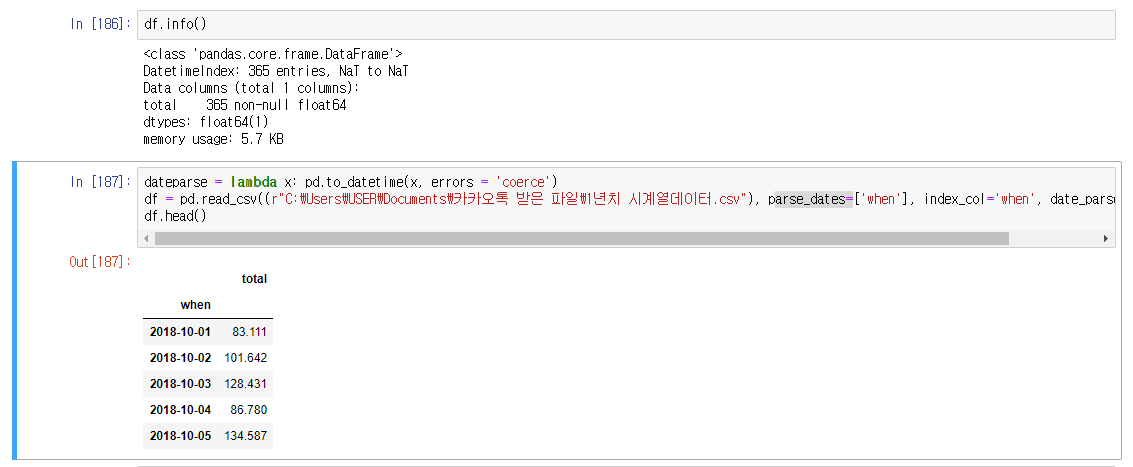
**<1년치 시계열데이터>**



1. 기본데이터 plot ( X축 : 날짜 , Y축 : 일일 총 배달건수 )



1. Info함수를 통해 (# of Data) = 365 임을 확인하고  
   dataparse 함수를 정의해서 데이터를 Datetime 형식으로 변경하였다.  
   또한 Index를 when column으로 변경해주었다.



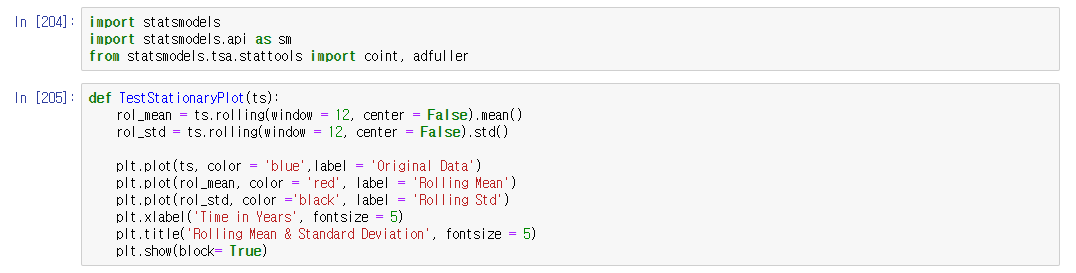
2번과 비교해보면, 변경 전 인덱스는 1 ~ 365, 변경 후 2018-10-01 의 형식으로 나타나는 것을 알 수 있다.

1. null값이 없는 데이터를 ts 변수에 대입 후 확인과정 
2. 시계열 정상성 확인과정

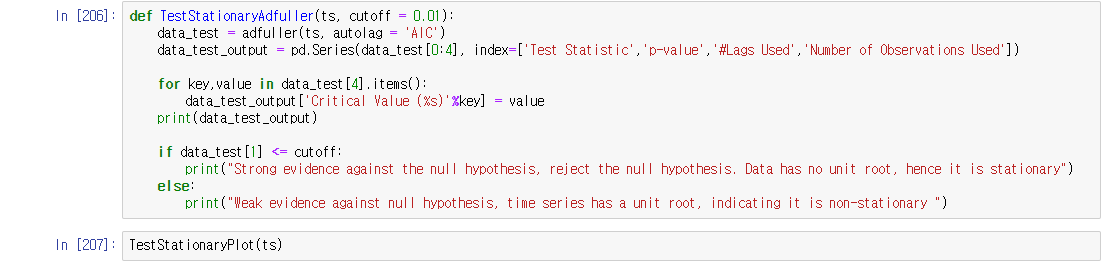
rolling함수를 이용해서 이동평균 및 분산을 이용해 정상성 확인

Window = 몇 개씩 연산할지를 결정한다.

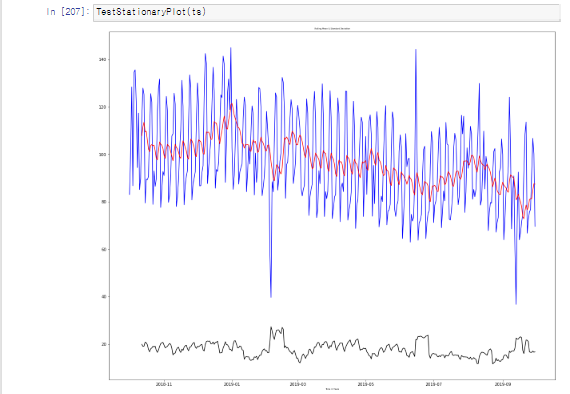
원데이터 : 파란색 / 이동평균 데이터 : 빨간색 / 이동분산 데이터 : 검은색

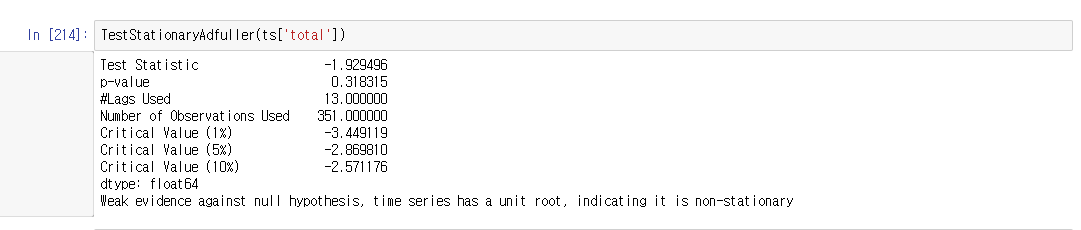


Dickey fuller 단위근 검정 : 적분 차수를 판단하기 위한 검정 방법

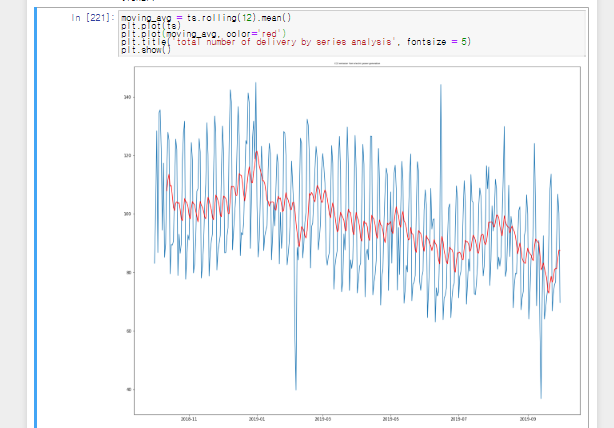


1. 이동평균 및 이동분산의 그래프

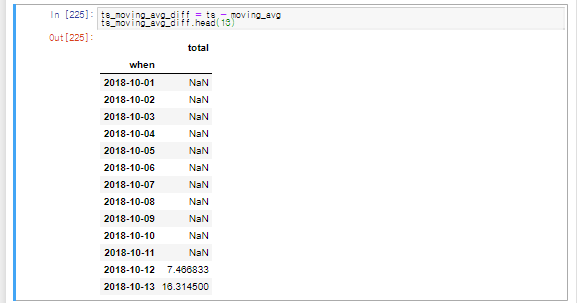


1. 90%, 95%, 99%가 p-value보다 작으므로 non-stationary
2. Dataset을 stationary로 변형

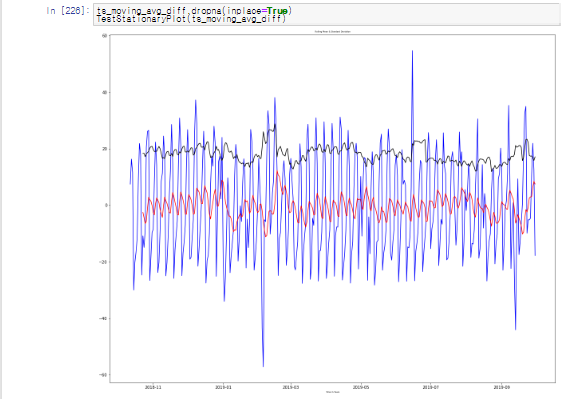
By 이동평균

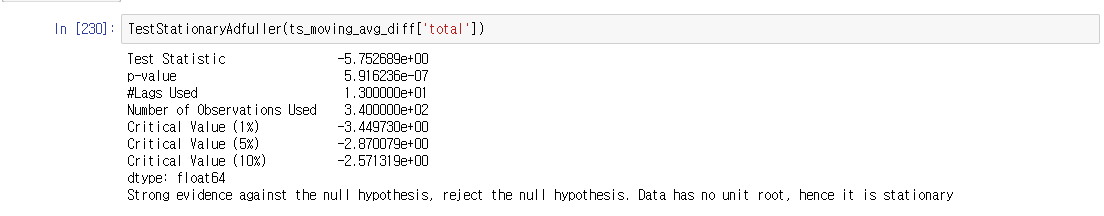
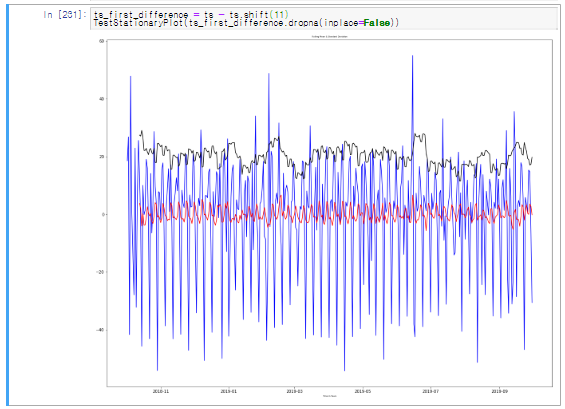


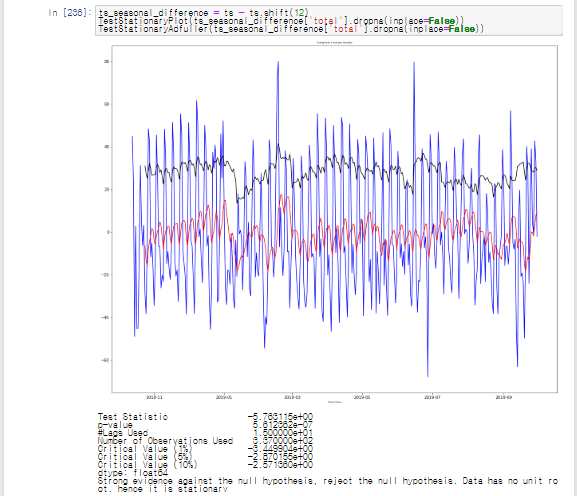
1. 이동평균이 window=12 때문에, 처음 11개의 데이터는 nan으로 나온다.

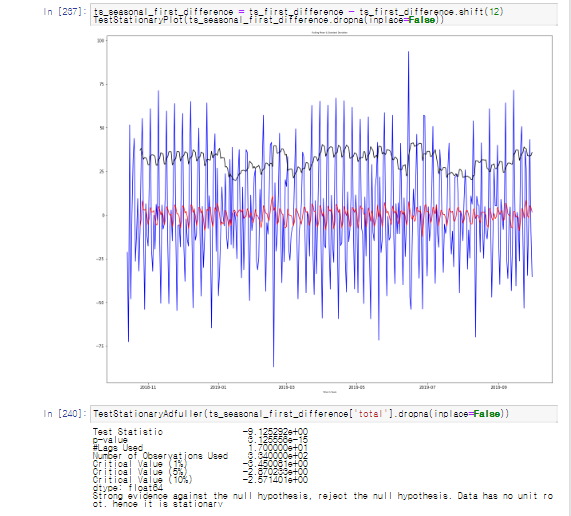


1. Nan 값을 버리고 다시 이동평균 및 이동분산을 구한다.

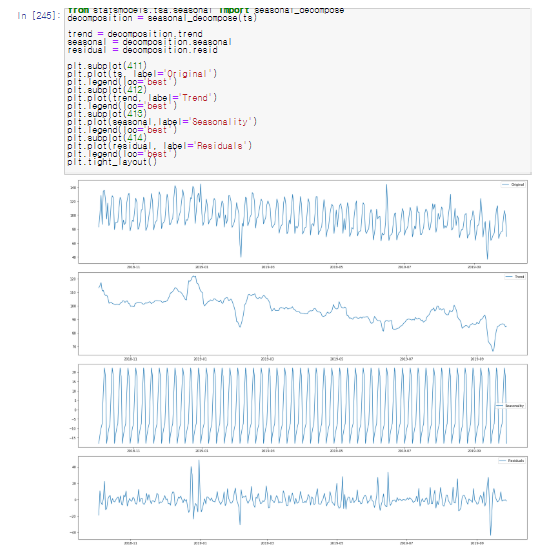


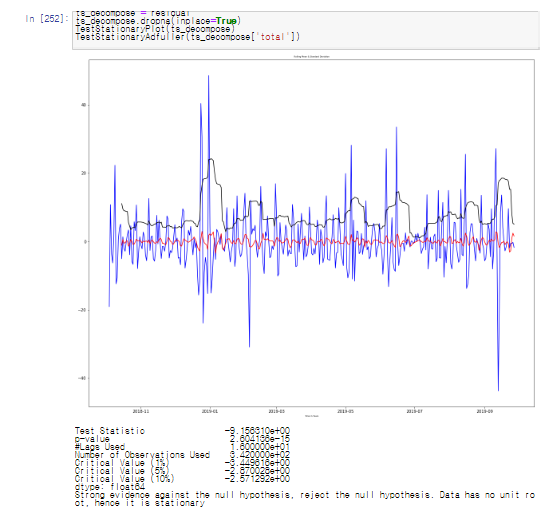
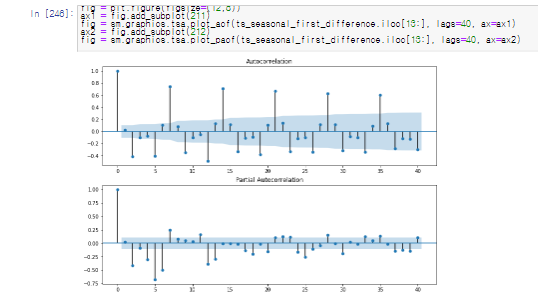
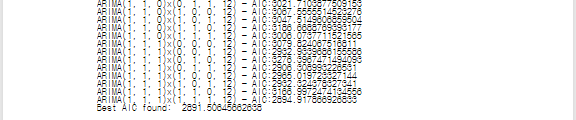
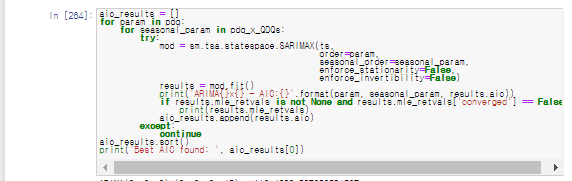
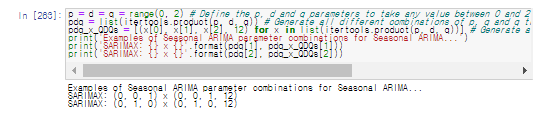
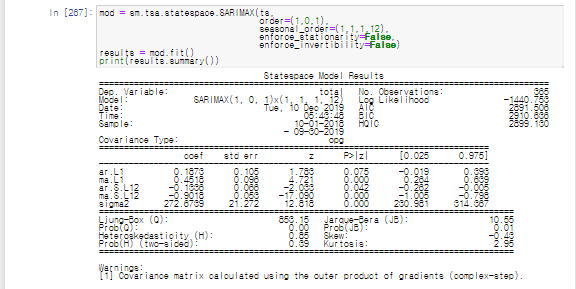
1. 이동평균은 약간씩 변화하고 있다는 것을 알 수 있다. 검정 통계량은 99%,95%,90%의 신뢰도는 p-value보다 작기 때문에 stationary이다. 
2. 추세와 계절성 제거 by 차분
3. Dickey fuller 단위근 검정 -> stationary로 변환 완료
4. 위와 같은 방법으로 계절성 또한 제거한다.

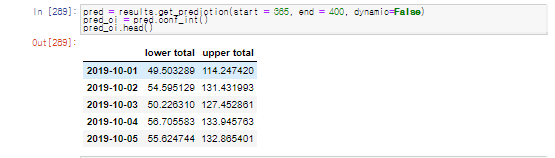
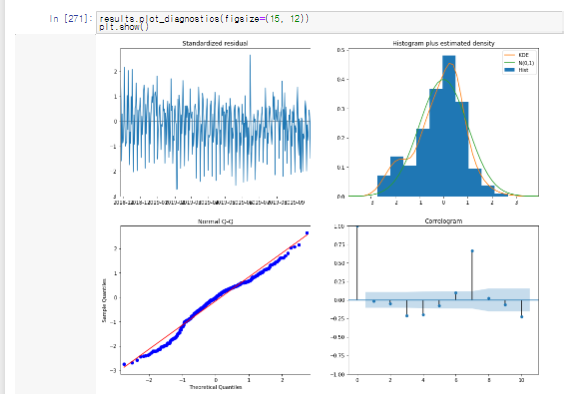
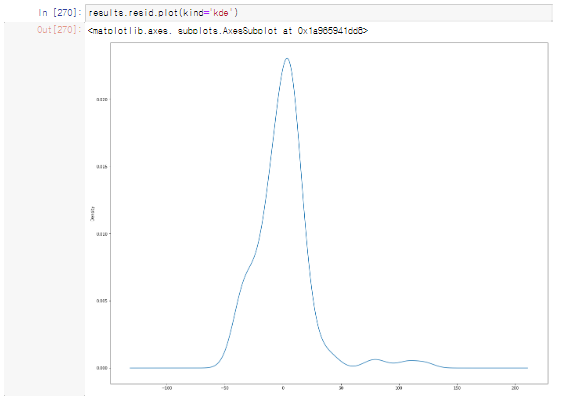
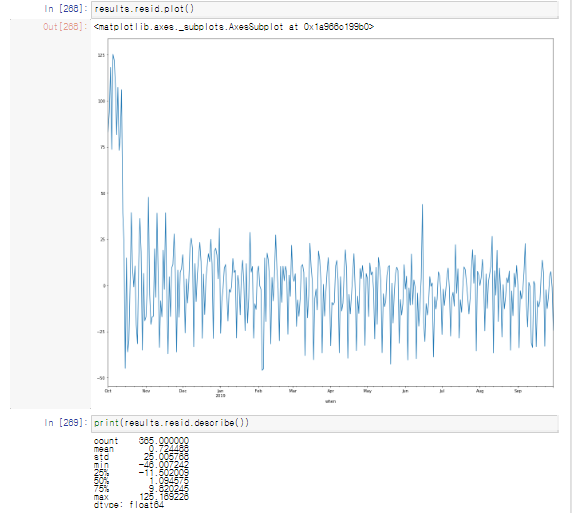


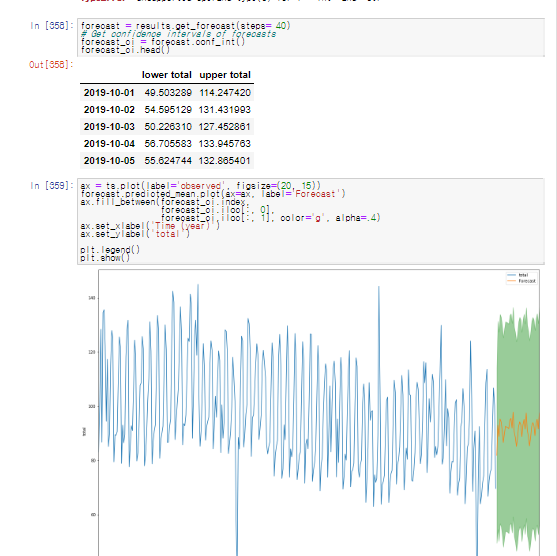
****

1. Decompose를 통해 트렌드와 계절성을 모두 모델링하고 모델에서 제거하여 통계화한다.



1. Residuals의 stationary를 체크 
2. ACF와 PACF 그래프
3. Grid search (최적해를 찾아주는 머신러닝 기법)   
   
4. SARIMA 결과



1. 예측결과 
2. **활용예시 및 기대효과**

자영업자들이 효율적인 마케팅을 하기위해 위 분석방법을 토대로 실제 데이터에 적용해 보았다.

기초 분석에 따르면 지점 중 도봉구의 총 배달 건수가 가장 낮기 때문에 도봉구의 배달건수 데이터를 이용하였다.

<도봉구의 **실제 치킨** 배달 건수>



<특정 날씨에 관한 도봉구의 **예측 치킨** 배달 건수>

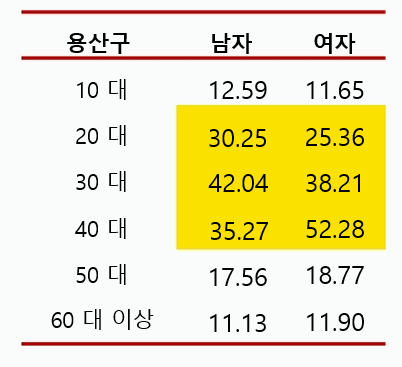


신경망 분석에서 사용한 구간화한 날씨 데이터를 사용하여 실제 활용 예시에 적용해 보았다. 날씨 데이터를 구간화 하였을 때 일반적으로 추울 것이라고 생각되어지는 데이터로 강수량 없음, 평균기온 구간 1, 평균풍속 구간 4, 일교차 구간 2의 특정 날씨 값을 적용하여 도봉구의 성별, 연령대 별 배달건수를 예측해 보았다.

위와 같은 날씨 값으로 예측해 본 결과 대체적으로 도봉구의 치킨 배달 건수는 감소하며 특히 40대의 배달 건수가 가장 많이 감소하는 것을 볼 수 있다. 따라서 도봉구는 예측되어진 배달 건수 값을 토대로 적은 평소보다 적은 인력배치를 해야 하고 40대는 남자와 여자의 배달 건수 값의 차이가 높기 때문에 해결방안을 모색해야 한다. 전체적으로는 배달건수가 줄었지만 남자보다 여자의 치킨 배달 건수가 많기 때문에 상대적으로 여성들이 좋아하는 치킨 메뉴의 식자재를 많이 준비해야 할 것이다.

기본분석의 시간대별 배달음식 주문 건수를 확인해보면 18시와 19시 사이에 가장 높은 판매량을 보이고 있으므로 도봉구의 치킨집 사장님은 이 시간대에 영업을 효율적으로 할 수 있도록 준비해놓아야 한다.

<같은 날씨의 용산구 예측 중국음식 배달건수>



같은 날씨의 용산구는 중국음식 배달 건수가 전체적으로 높고 그 중 20, 30, 40대가 높은 것을 알 수 있다. 즉 추운 날씨에 용산구는 20,30,40대를 중점으로 마케팅 해야한다. 중국음식은 점심시간대인 11시부터 13시와 저녁시간대인 17시부터 19시 모두 배달 건수가 많으므로 용산구의 중국음식 자영업자는 이를 토대로 20, 30, 40대가 모두 즐겨 먹을만한 메뉴를 준비하고 영업이 유연하게 잘 돌아가도록 인력배치를 많이 해야 함을 알 수 있다.

위의 활용예시처럼 각 지점의 자영업자들은 기상청에서 예측된 날씨데이터를 모델에 입력하면 그 날씨에 예측되어지는 주문건수를 알 수 있으므로 특정 날씨에 어떻게 인력 배치를 해야하고 식자재를 준비해야 하는지 알 수 있다. 또한 특정 날과 시간대에 연령대마다 선호하는 음식을 알 수 있기 때문에 연령대에 따른 알맞은 마케팅 전략을 세울 수 있다. 또한 새롭게 배달업체 시장에 들어올 신규 사업자들은 이 모델을 사용하여 어느 지점에 어떤 음식의 배달건수가 높은지 알 수 있으므로 어떤 지점에 점포를 내야 이익을 얻을 수 있는지 알 수 있기 때문에 영업에 한층 더 도움이 될 수 있다.

즉 날씨에따른 치킨, 피자, 중국음식 주문량을 예측하여 자영업자들이 활용하기 어려웠던 날씨 데이터를 손쉽게 활용하여 정보를 제공할 수 있다. 또한, 식자재의 효율적인 운용과 특정 날과 시간대에 따른 소비 고객 군의 변화를 파악하여 효율적인 마케팅 전략을 수립할 수 있을 것이며 여러가지 전략을 제안함으로써 사업 활성화에 기여할 수 있을 것이다.

1. **느낀점**

spss modeler를 처음 배운 입장으로서 처음에는 분석의 품질보다는 분석기법의 양에 초점을 맞추었다. 그래서 처음에는 군집화와 회귀분석 등 최종분석에 사용하지않았던 것에 시간을 투자했다. 목적에 맞는 분석이 아닌 그냥 다양한 분석기법으로 좋은 결과를 내려고 노력하니 당연히 그러지 못하였다. 또한, R이나 파이썬과 다르게 코드를 직접 써서 이용하는 것이 아니였기 때문에 편한 줄만 알았으나 설정하는 것 하나하나의 의미에 따라 조금씩 변하는 결과를 보고 오히려 더욱 어렵다고 느낄 때도 있었다. 적절한 데이터로 적절한 분석기법을 사용하는 것이 얼마나 어렵고 까다로운 일인지 배웠으며 tool은 사람과 같이 응용 및 생각하는 능력이 없어 잘못입력을 해도 그대로 받아들여 만족스러운 결과값이 나오지 않는 경우가 많았음을 알게되었다. 분석 기법으로 인해 나온 결과값으로 인해 타당한 결론을 도출하는 것도 통계적 기반이나 지식이 없으면 그저 복잡한 수치에 불가한 것임을 깨달았다. 데이터들을 올바르게 얻어 분석만 잘 하면 될 것 같았지만 우리의 텀프로젝트는 분석 하나하나가 고됬고, 특히 무엇보다도 시계열 분석이 극강의 난이도를 보이고 있었다. 솔직히 시계열의 개념도 구축하지 않고 바로 코딩 및 분석을 하려고 보니 수차례 기본 정의나 단어,개념을 찾아봐야 했고 시계열의 토대가 되는 확률 및 통계에 대한 공부가 바탕이 되어야한다고 생각했다. 시계열 분석을 modeler, R, Python 3가지 tool로 돌렸으나 어느 하나 우리 팀 목적에 맞는 제대로 된 결과가 나오지 않아 포기하고 싶었지만, 어떻게든 해내고싶은 마음에 데이터를 1년치에서 4년치로 확대하여 분석한 결과 만족할 만한 결과가 나온 것 같다. 시계열 분석이 올바르게 되지 않았던 이유를 참고자료나 서적에서 찾아본 결과, 주기의 문제라고 파악되나 확실한 문제는 찾아내지 못하였다. 기회가 된다면 통계학과 지식을 공부한 뒤 다시한번 올바른 분석을 해보고 싶다.