1. 데이터셋 개요

2. 전처리 및 특징 분석

3. 모델 서칭 및 실험 방법

4. 성능 및 실행시간

5. 결과 분석

6. 결론

1. 데이터셋 개요

Tourism : 월별, 분기별, 연간 관광객

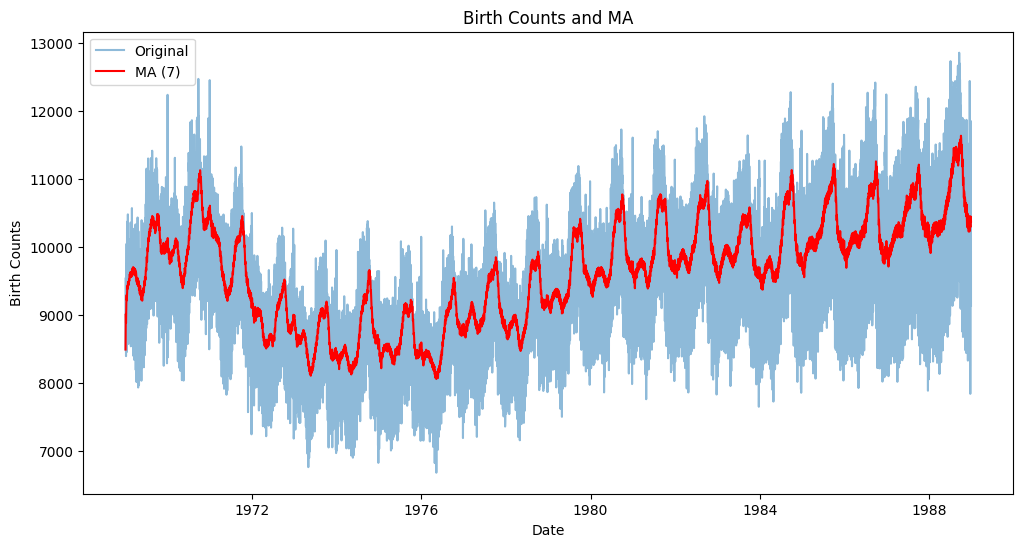
Rideshare : 2018년 11월 26일~2018년 12월 18일 뉴욕의 다양한 지역에서 Uber 및 Lyft 승차 공유 서비스와 관련된 속성의 다양한 시간별 시계열

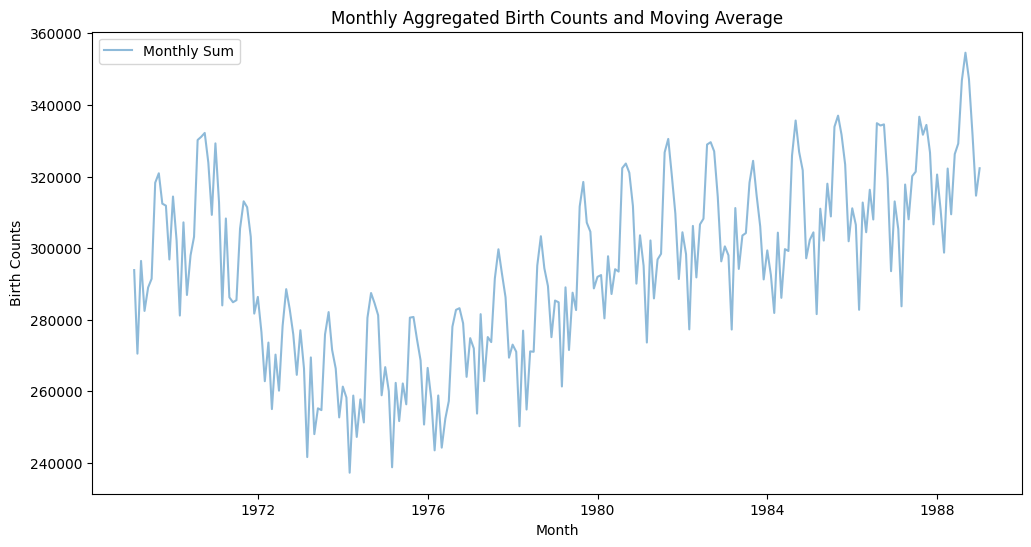
Sunspot : 1818년~2020년 일별 흑점데이터 세트

US birth : 1969년~1988년 일별 미국에서 태어난 아기 수

1. 전처리 및 특징 분석

US birth, 1969년, 1988년 일별 태어난 아기 수. 대략적으로도 패턴과 추세가 관측됨





월간 리샘플 시계열

ACF와 KPSS결과가 상반되어 1차 차분. 차분후 ADF, KPSS검정 값은 Stationary.

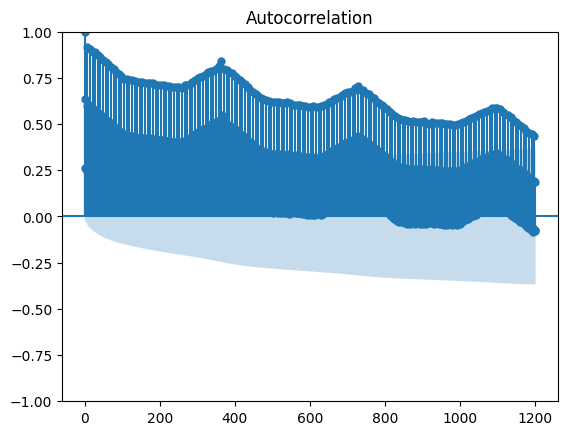
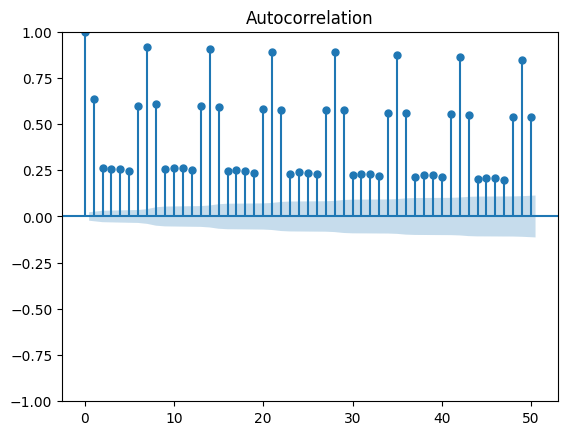
Sunspot과 마찬가지로 원 데이터가 정상과 비정상의 경계에 있는듯(약한 추세가 관측되어 그런거같음)

텍스트, 스크린샷, 그래프, 라인이(가) 표시된 사진

자동 생성된 설명

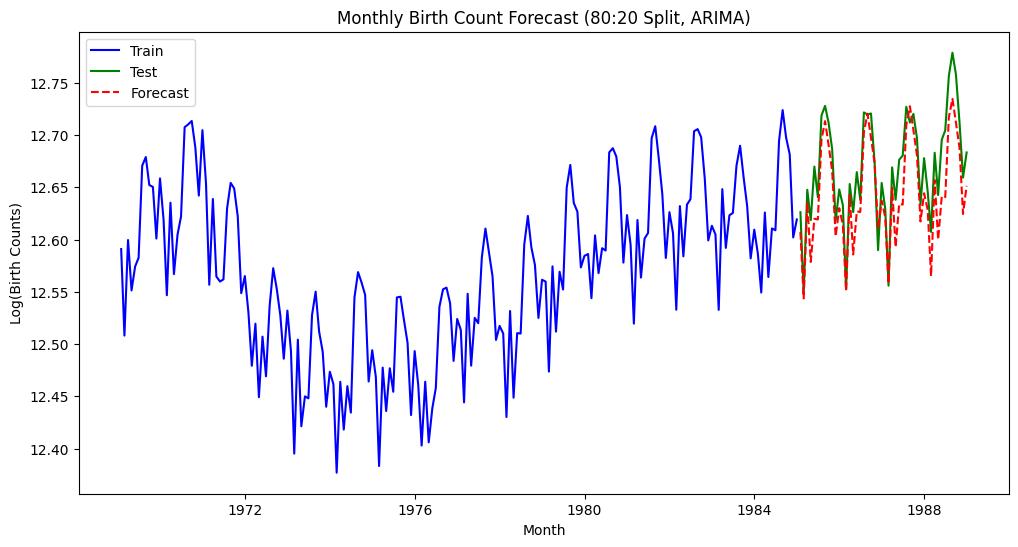
특이하게도, 약 7일마다 단기적인 패턴이 관측. -> 주간 계절성

약 1년마다 연간 계절성도 관측. 즉 장단기 주기를 모두 가지고 있음



추세, 연간 계절성을 모두 고려해 SARIMA 사용.

주간 계절성은 학습 편의를 위해 일단 스킵. 원본 데이터를 사용하면 학습 시간이 너무 오래걸려 월간 리샘플(일별 값을 다 더함) 후 학습.



벤치마크 평가치(MASE)

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

실제 모델 평가치(Old and New)

new

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

old

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

New가 벤치마크보다 더 잘 나온 이유? -> 데이터를 리샘플해 노이즈가 줄어 패턴이 훨씬 러프하게 관측되었기 때문인듯.

Old 데이터는 그냥 참고로 넣은 것으로, 대충 처음 돌렸을 때 나온 값임. 가장 최근 값과 통제되지 않은 조건에서 실행한 값이므로 실행시간 말고 값 자체는 의미 없음. 참고만