2021년 공공 빅데이터 청년인턴십 결과보고서

선박 충돌 예방을 위한
해양사고 현황분석

2021.02.25

21-13230	이준형

인천지방해양안전심판원

목 차

1. 분석 개요

1-1. 배경 및 필요성

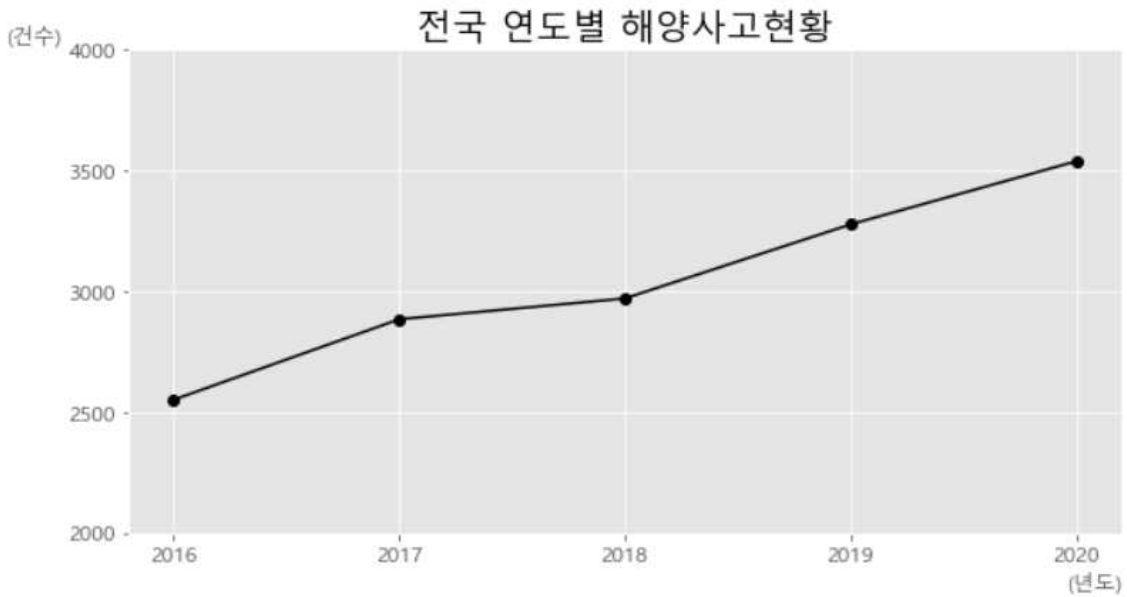
□ 배경

이미 현존하는 수많은 해양안전대책에도 불구하고 우리나라에서 해양사고는 계속해서 발생하고 있다. 이는 주체자인 어민들과 행정 실무자들 사이에서 발생하는 이해관계 차이에 비롯한 문제로 보인다.

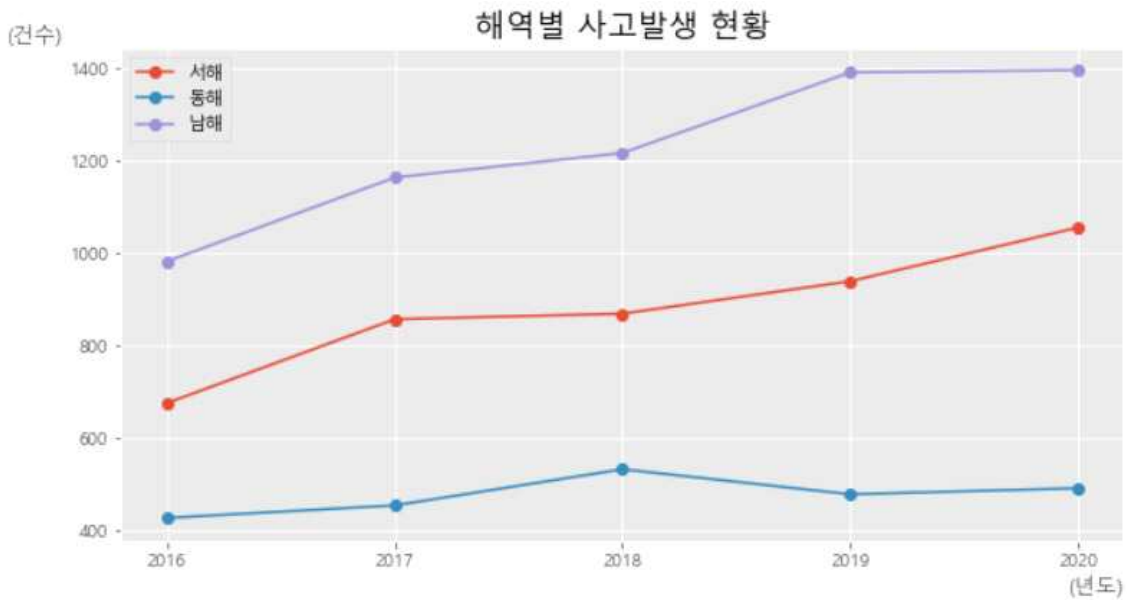
따라서 새로운 방법만을 제시하는 것이 아닌 기존의 대책들이 보다 실질적으로 이루어 질 수 있도록 어민들의 안전관리 엄수를 촉구하는 근거자료가 필요하다.

[그림 1-1] 해양사고 예방대책 강구 (기사제목)

□ 해양사고 현황

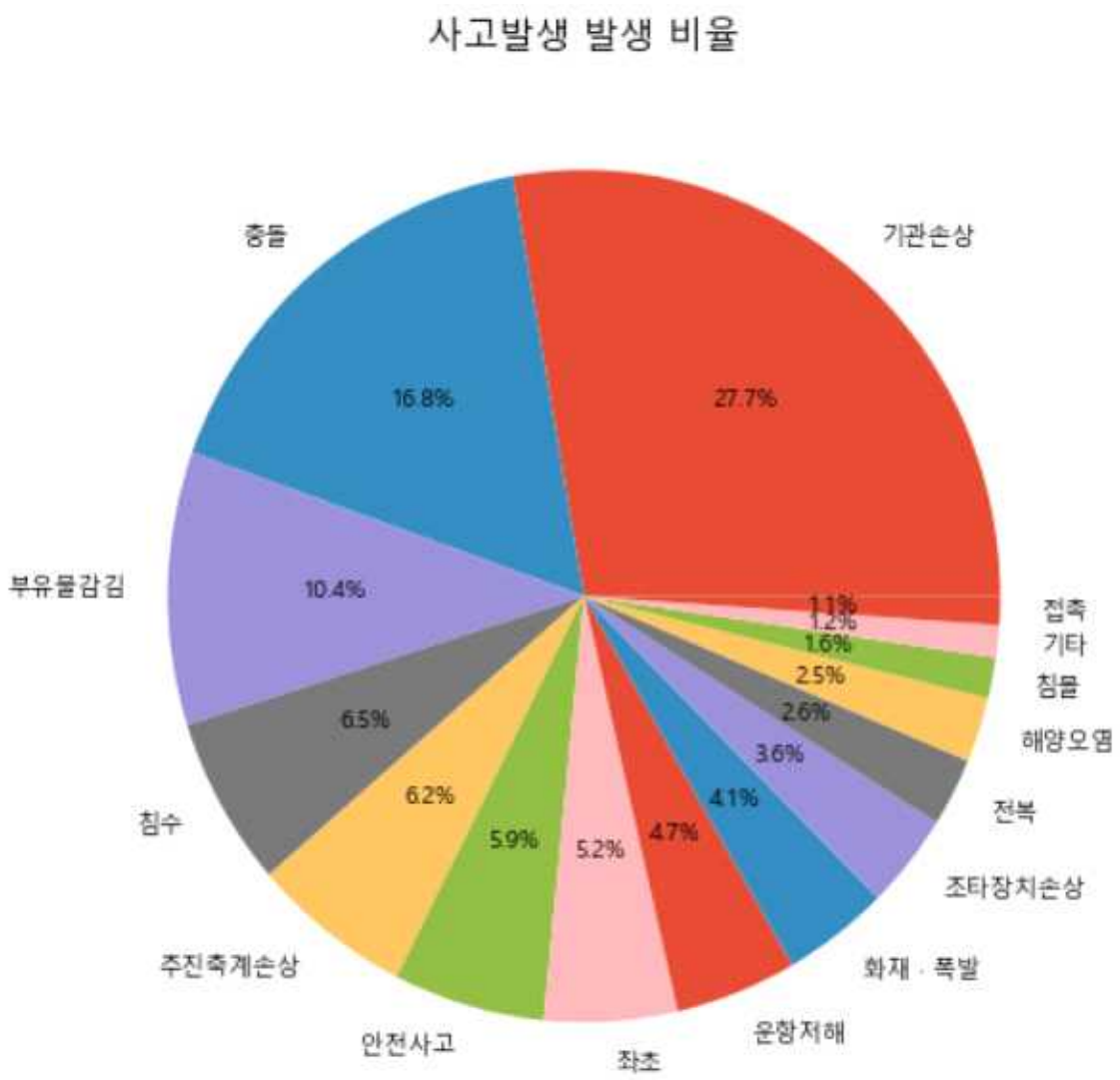


[그림 1-2] 전국 연도별 해양사고 발생 현황 (2016~2020)



[그림 1-3] 해역별 사고발생 현황 (2016~2020)
(*서해, 동해, 남해는 각각 영해와 공해를 모두 합친 해역을 의미한다)

2016년~2020년도까지의 해양사고 발생률은 점차 증가하는 추세에 직면해 있으며 그 중 '인천지방해양안전심판원'의 관할구역인 서해안의 해양사고율은 현저히 증가하는 추세이다.

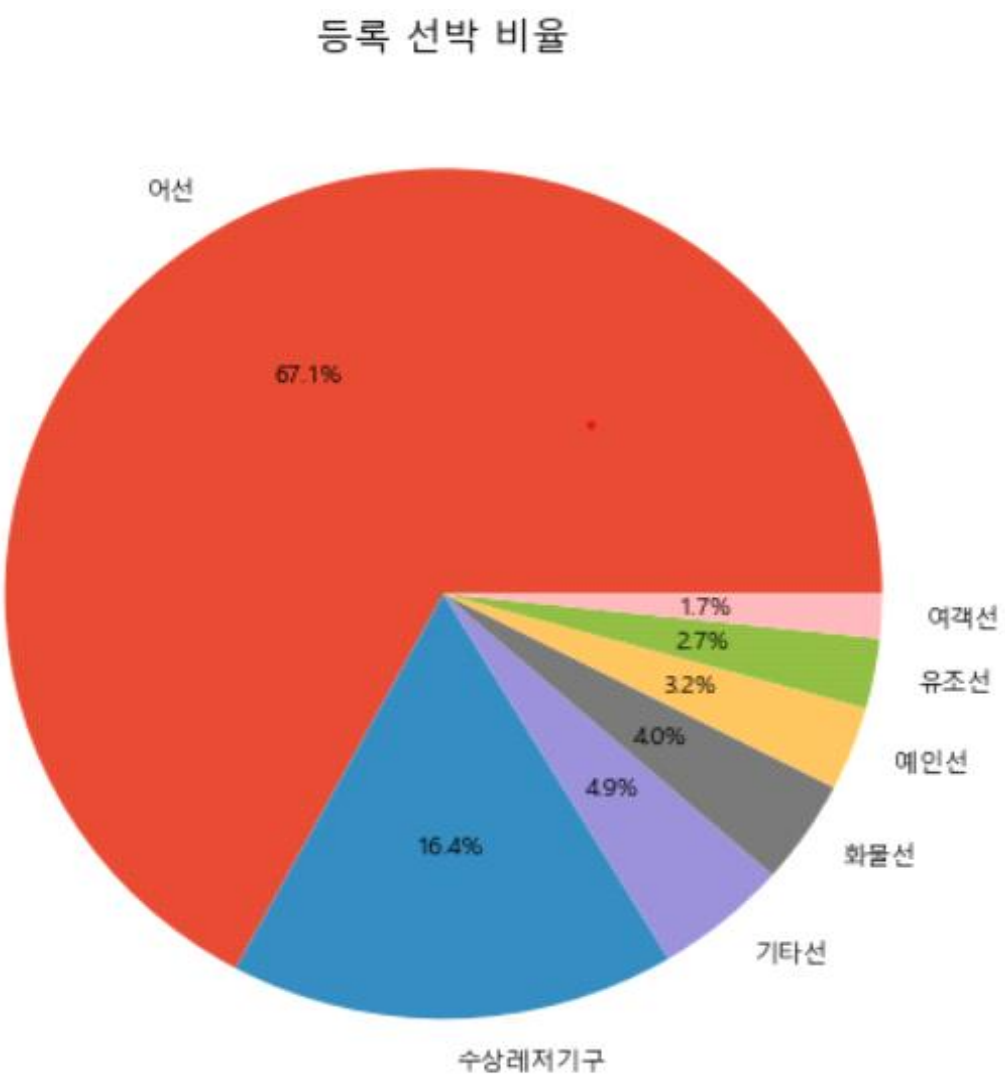


[그림 1-4] 해양사고 발생 비율 (2016~2020)

전국 해양사고 발생 빈도 상위 50%사건들의 종류들을 살펴보았을 때 '기관손상 27.7%, '충돌' 16.8%, '부유물 감김' 10.4%로 나타났다.

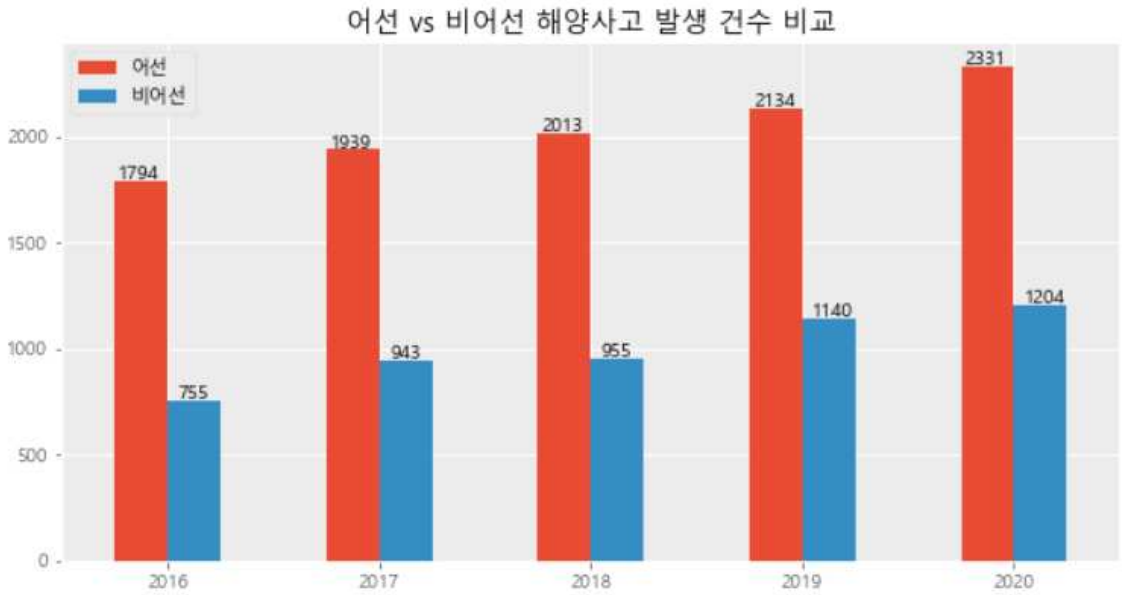
'기관손상'과 '부유물 감김'은 원인을 특정할 수 없고 비교적 경미한 사건으로 주로 심판이 청구되지 않는 '심판불필요처분'에 해당한다.

따라서 보다 위험성이 높고 현실적 예방책을 제시할 수 있는 '충돌사건'의 현황을 분석하는 것을 목표로 하며, 나아가 시계열 예측 분석을 통해 해당 실무자들에게 현 상황을 직선적으로 바라볼 수 있는 지표를 제시한다.



[그림 1-6] 등록 선박 비율 (2016~2020)

등록선박의 비율로 어선이 무려 67.1%를 차지할 정도로 대부분을 차지하고 있으며, 조업 가능한 연해 및 근해에서 고루 분포하고 있는 만큼 '어선'들의 충돌 예방 안전관리에만 성공하더라도 충돌사고 감소에 효과적일 것이라 기대된다.



[그림 1-7] 어선 vs 비어선 해양사고 발생 건수 비교 (2016~2020)

통계적 그래프로 확인해보아도 어선의 해양사고 발생 비율이 비어선에 비해 월등히 많은 것을 알 수 있다.

[그림 1-8] 어선의 해양사고 종류별 비율

'기관손상'과 '부유물 감김'을 제외하고 어선은 가장 높은 비율로 '충돌사고'(16.4%)가

발생하는 것으로 나타났다. 따라서 주요 해양사고 발생요소인 '어선'을 특정하고 어선의 '충돌'이 감소한다면 향후 해양 충돌사고 발생률 또한 감소하는 추세로 접어들 것이다.

□ 기대효과 및 필요성

사고 예방을 위한 선제적 대응이 중시되는 현 시점에서, 본 분석 프로젝트는 어선들의 안전관리 촉구 및 사고 예방 대책 제시에 근거를 마련하는 것을 목표로 하며, 어민들과 행정 실무자들 사이 이해관계의 간극을 좁히고 실질적인 대응책 마련에 효과적일 것으로 기대한다.

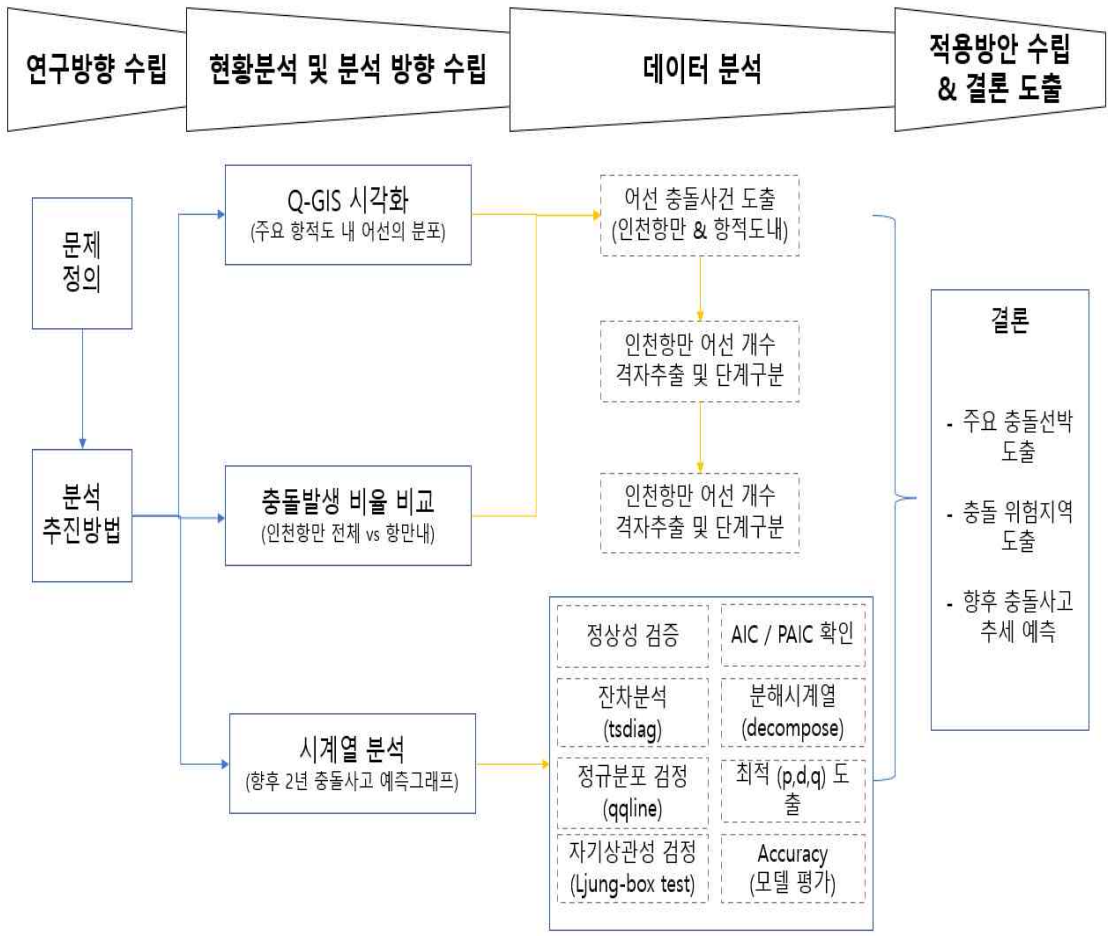
1-2. 수행기간

[그림 1-8] 분석 수행기간

1-3. 분석 목표

- '충돌사건'이 잦은 주요 위험지역 도출
- 인천항만 충돌사건 현황분석 후 주요 선박 항로와의 상관성 파악
- 시계열 분석을 통해 사고의 추세, 계절성, 주기성의 존재유무 확인
- 시계열 분석으로 실무 적용을 위한 향후 미래 예측

1-4. 수행체계

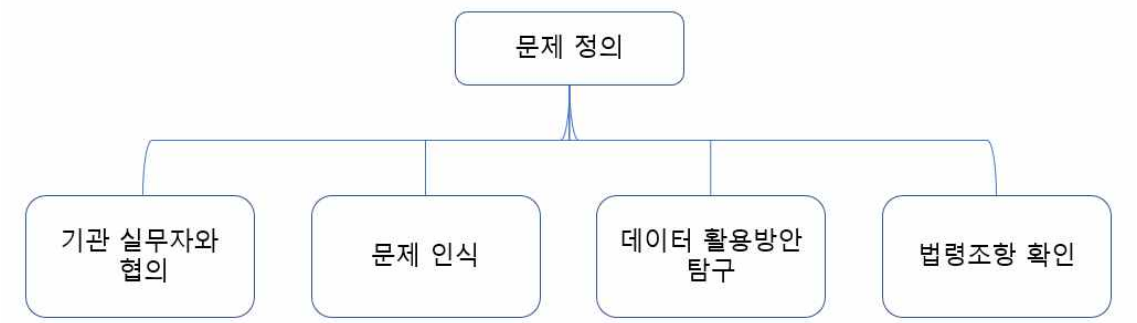


[그림 1-9] 분석 수행체계

2. 분석 방법

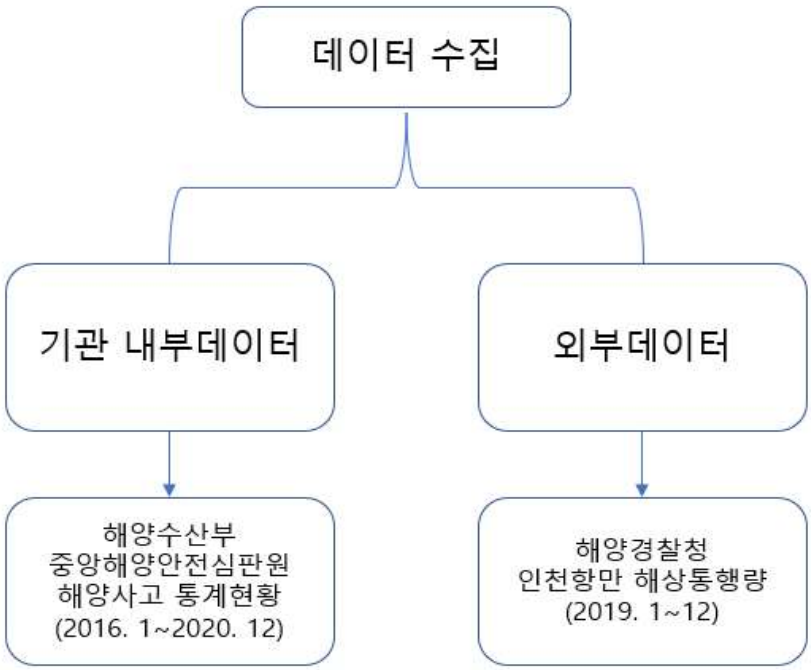
2-1. 분석 프로세스

(1) 문제 정의



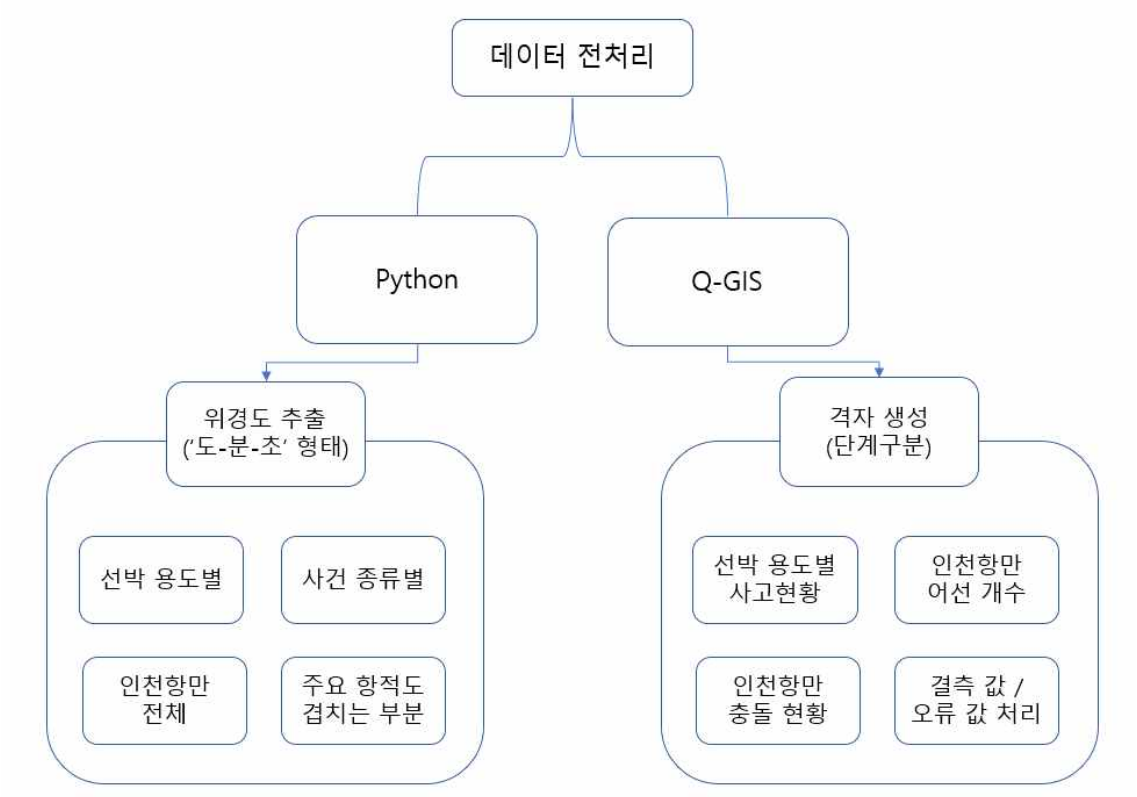
[그림 2-1] 문제정의

(2) 데이터 수집



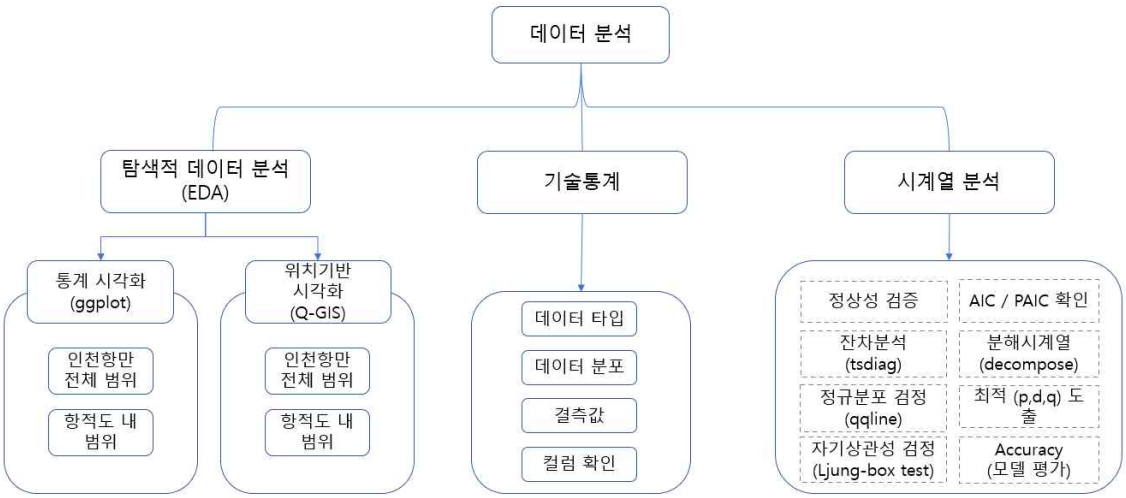
[그림 2-2] 데이터 수집

(3) 데이터 처리



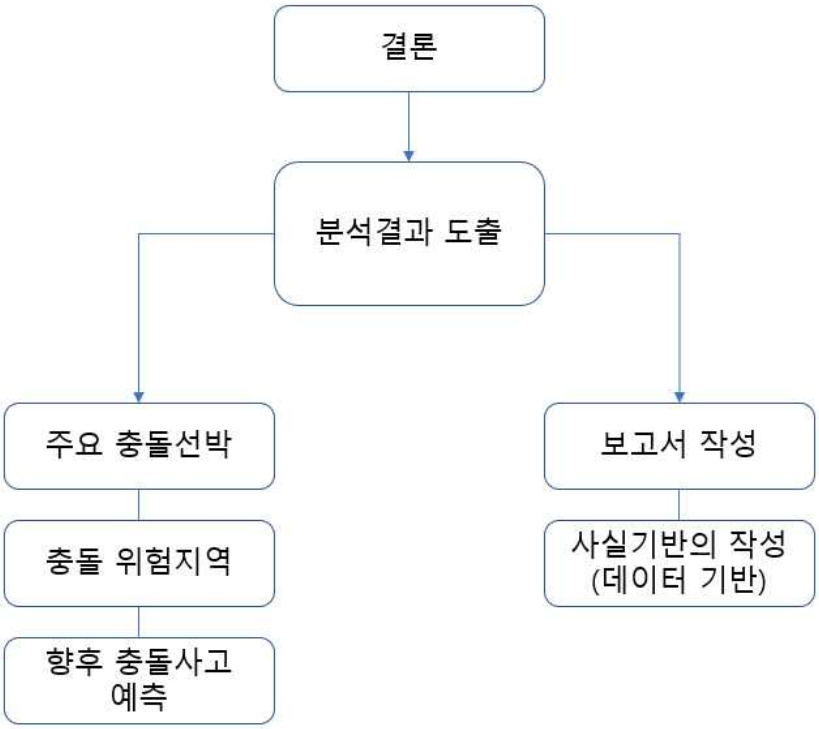
[그림 2-3] 데이터 전처리

(4) 데이터 분석



[그림 2-4] 데이터 분석

(5) 결론



[그림 2-5] 결론도출

2-2. 분석 도구 / 환경

R studio

[시계열 분석]

: 정상성 검증, 잔치 분석, 정규분포 검정, 자기상관성 검정, AIC/PAIC 확인, 분해시계열, auto.arima, Accuracy

python

[데이터 전처리 / EDA]

: 기술통계, 결측/오류 처리, 컬럼추출, 데이터 구조화, ,EDA시각화

QGIS

[공간시각화]

: 그리드 생성, 격자추출, 주요 사고발생지역 도출, 오류 위경도 정보 제거

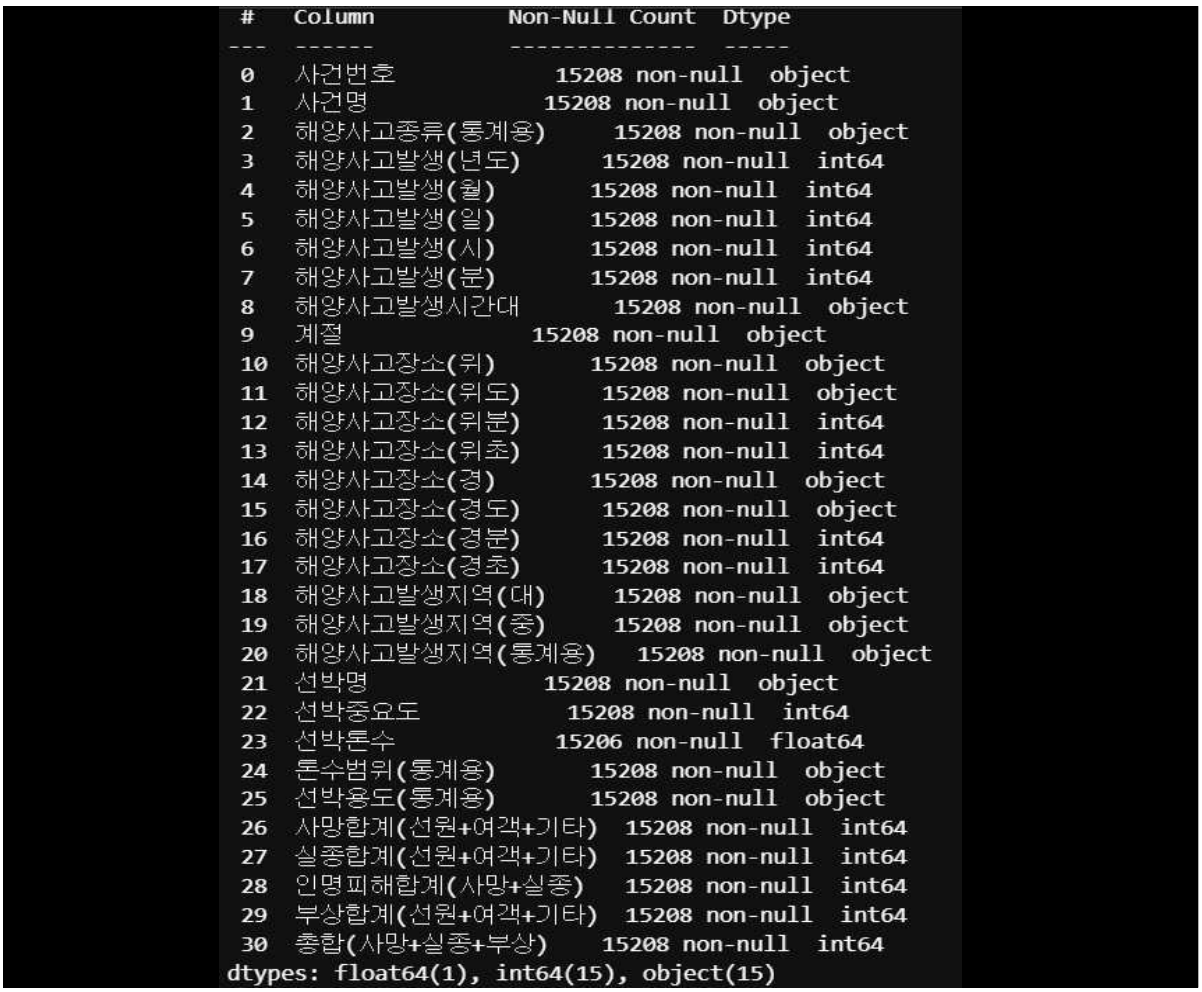
2-3. 활용 데이터

데이터명	구분	파일형식	컬럼명
해양수산부 중앙해양안전심판원_해양사고 통계현황_20200331	내부데이터	Microsoft Excel 쉼표로 구분된 값 파일(.csv)	사건번호
			사건명
			해양사고종류(통계용)
			해양사고발생(년도)
			해양사고발생(월)
			해양사고발생(일
			해양사고발생(시)
			해양사고발생(분)
			해양사고발생시간대
			계절
			해양사고장소(위)
			해양사고장소(위도)
			해양사고장소(위분)
			해양사고장소(위초)
			해양사고장소(경)
			해양사고장소(경도)
			해양사고장소(경분)
			해양사고장소(경초)
			해양사고발생지역(대)
			해양사고발생지역(중)
			해양사고발생지역(통계용)
			선박명
			선박중요도
			선박톤수
			톤수범위(통계용)
			선박용도(통계용)
			사망합계(선원+여객+기타)
			실종합계(선원+여객+기타)
			인명피해합계(사망+실종)
			부상합계(선원+여객+기타)
			총합(사망+실종+부상)
해양경찰청_ 인천항만 해상통행량_20191231	외부데이터	SHP 파일(.shp) PRJ 파일(.prj) SHX 파일(.shx) CPG 파일(.cpg) DBF 파일(.dbf)	ID
			LEFT
			RIGHT
			TOP
			BOTTOM
			TRAFFIC

[표 2-1] 활용데이터

2-4. 분석 방법

□ 기술통계



```
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   사건번호                15208 non-null  object
 1   사건명                 15208 non-null  object
 2   해양사고종류(통계용)       15208 non-null  object
 3   해양사고발생(년도)        15208 non-null  int64
 4   해양사고발생(월)         15208 non-null  int64
 5   해양사고발생(일)         15208 non-null  int64
 6   해양사고발생(시)         15208 non-null  int64
 7   해양사고발생(분)         15208 non-null  int64
 8   해양사고발생시간대         15208 non-null  object
 9   계절                  15208 non-null  object
 10  해양사고장소(위)         15208 non-null  object
 11  해양사고장소(위도)        15208 non-null  object
 12  해양사고장소(위분)        15208 non-null  int64
 13  해양사고장소(위초)        15208 non-null  int64
 14  해양사고장소(경)         15208 non-null  object
 15  해양사고장소(경도)        15208 non-null  object
 16  해양사고장소(경분)        15208 non-null  int64
 17  해양사고장소(경초)        15208 non-null  int64
 18  해양사고발생지역(대)       15208 non-null  object
 19  해양사고발생지역(중)       15208 non-null  object
 20  해양사고발생지역(통계용)     15208 non-null  object
 21  선박명                 15208 non-null  object
 22  선박중요도              15208 non-null  int64
 23  선박톤수               15206 non-null  float64
 24  톤수범위(통계용)         15208 non-null  object
 25  선박용도(통계용)         15208 non-null  object
 26  사망합계(선원+여객+기타)  15208 non-null  int64
 27  실종합계(선원+여객+기타)  15208 non-null  int64
 28  인명피해합계(사망+실종)   15208 non-null  int64
 29  부상합계(선원+여객+기타)  15208 non-null  int64
 30  총합(사망+실종+부상)     15208 non-null  int64
dtypes: float64(1), int64(15), object(15)
```

[그림 2-6] '해양수산부 중앙해양안전심판원_해양사고 통계현황_20200331' .info()

주요 내부데이터인 '해양수산부 중앙해양안전심판원_해양사고 통계현황_20200331'은 2016.01 ~ 2020.12 까지 해양사고 접수현황 데이터이다.

총 31개의 컬럼, 15208개의 열로 구성된 데이터 형태이며 '선박용도' 컬럼에서 2개의 결측값이 확인된다.

	해양사고장소(위)	해양사고장소(위도)	해양사고장소(위분)	해양사고장소(위초)	해양사고장소(경)	해양사고장소(경도)	해양사고장소(경분)	해양사고장소(경초)
0	북	NN36	41	0	동	동E126	7	0
1	북	NN30	51	0	동	동E126	58	0
2	북	NN34	53	6	동	동E126	23	13
3	북	NN37	6	15	동	동E126	40	48
4	북	NN36	24	40	동	동E126	21	18
...	...	...	...	...	...	...	...	...
15203	북	NN29	17	0	동	동E125	35	0
15204	북	NN35	2	5	동	동E129	4	58
15205	북	NN35	5	7	동	동E129	0	14
15206	북	N35	3	54	동	E129	0	41
15207	북	N35	5	0	동	E129	43	0

[데이터 전처리 전]

	사건명	해양사고종류(통계용)	위도	경도
0	레저보트 거북호 침수사건	침수	36.683333	126.116667
1	어선 제212영남호, 중국어선 선명미상 충돌사건	충돌	30.850000	126.966667
2	어선 만성호 무등록선박 예천호 충돌사건	충돌	34.885000	126.386944
3	어선 선명무 운항저해사건	운항저해	37.104167	126.680000
4	레저보트 서해킹호 기관손상사건	기관손상	36.411111	126.355000
...	...	...	...	...
15203	어선 제37진성호, 컨테이너운반선 절보어운68987 충돌사건	충돌	29.283333	125.583333
15204	원양어선 신유한호 해양오염사건	해양오염	35.034722	129.082778
15205	원양어선 카피딴 팰리예브(KAPITAN FALEYEV) 해양오염사건	해양오염	35.085278	129.003889
15206	어선 카피탄마슬로베츠.어획물운반선 헬시8 충돌사건	충돌	35.065000	129.011389
15207	원양어선 세종 기관손상사건	기관손상	35.083333	129.716667

15208 rows × 4 columns

[데이터 전처리 후]

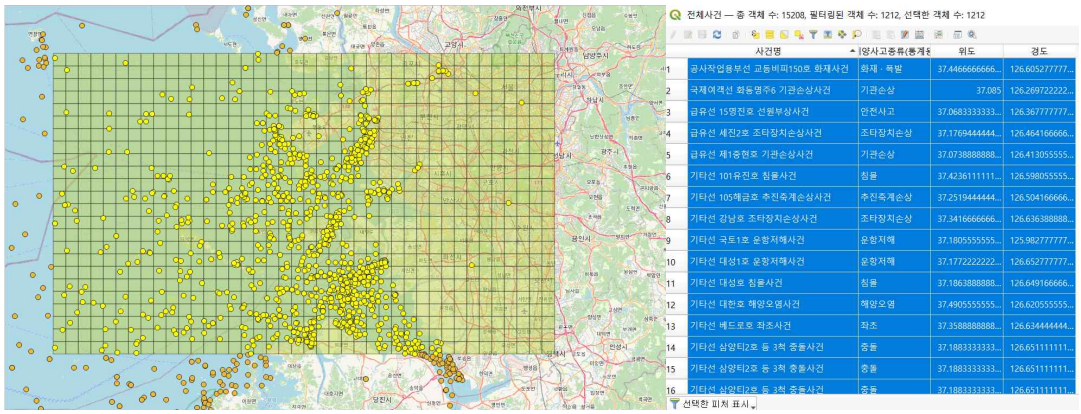
[그림 2-7] 위경도 추출

GIS 지도상에 공간시각화를 위해서는 '도-분-초'형태가 아닌 '위도', '경도' 형태의 데이터 컬럼을 필요로 하므로 적절한 형태로의 변환과정을 거친다.

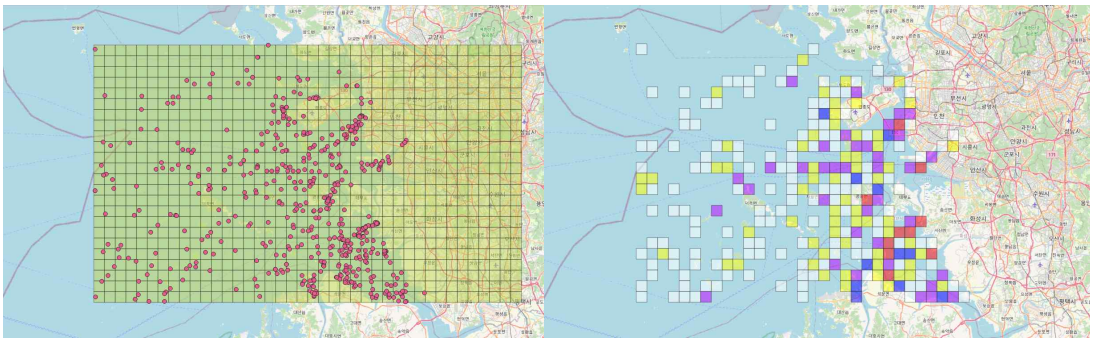
해양사고종류별로 추출한 위경도 데이터를 통해 사건의 종류별 발생 현황을 GIS에서 시각화하고 이에 따른 인사이트를 도출 할 수 있다.

□ Q-GIS 시각화

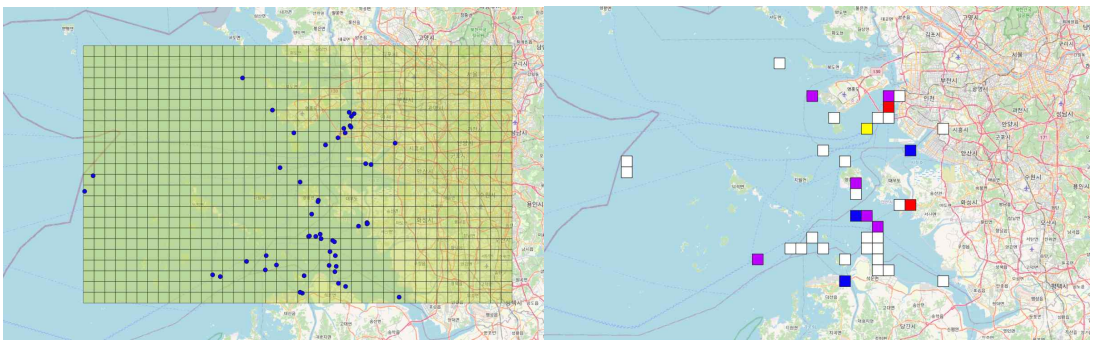
1. 인천항만 내 사건 특정



2. 인천항만 사건들 중 '어선'의 해양사고 분포 시각화



3. 인천항만 내 '어선'의 '충돌사건' 시각화



[그림 2-8] GIS시각화 프로세스

인천항만 내 어선사고의 분포 현황과 주요 충돌사고 위험지역을 도출하였다. 주 항로의 가변성이 낮은 대형선박들(화물선, 유조선, 여객선 등)의 항적도와 비교(대치)하여 '충돌'에 있어서 어선분포의 위험성을 직시할 수 있고, 이에 따라 안전관리 엄수를 촉구하는 근거를 마련할 수 있다.

□ 시계열 분석

[전처리 및 EDA]

1. 기존의 시간정보들을 datetime형태로 변환

	해양사고발생(년도)	해양사고발생(월)	해양사고발생(일)	해양사고발생(시)	해양사고발생(분)	해양사고발생시간대
0	2016	5	12	15	20	12-16시
1	2016	3	22	0	0	0-4시
2	2016	11	12	20	54	20-24시
3	2017	5	1	6	28	4-8시
4	2016	5	22	18	12	16-20시
...	...	...	...	...	...	...
15203	2017	11	8	14	0	12-16시
15204	2017	5	5	8	33	8-12시
15205	2017	11	9	9	30	8-12시
15206	2019	7	29	11	14	8-12시
15207	2020	12	17	21	7	20-24시

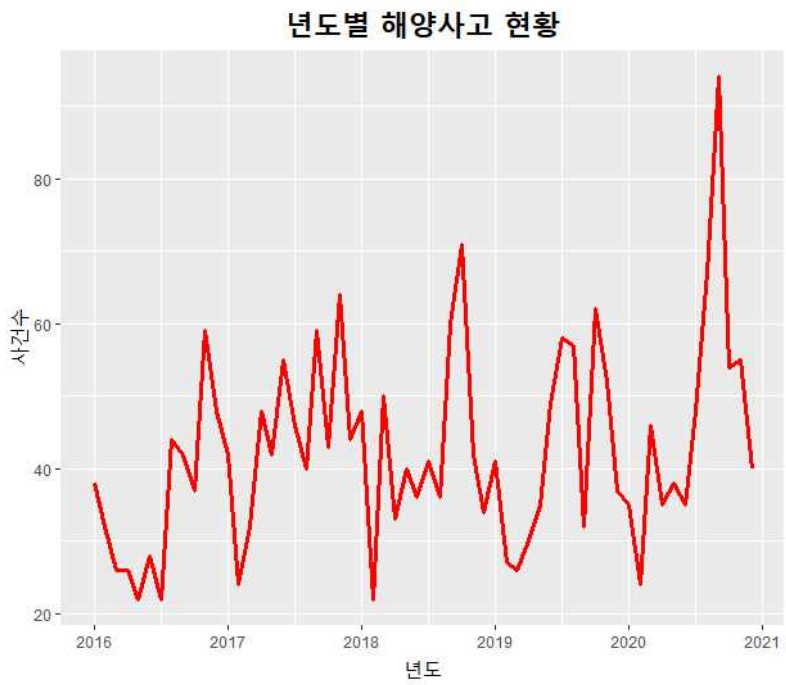
	사건명	해양사고발생시간(월별)
0	레저보트 기북호 침수사건	2016-05-01
1	어선 제212영남호, 중국어선 선명미상 충돌사건	2016-03-01
2	어선 만성호 무등록선박 예천호 충돌사건	2016-11-01
3	어선 선명무 운항저해사건	2017-05-01
4	레저보트 서해킹호 기관손상사건	2016-05-01
...	...	...
15203	어선 제37진성호, 컨테이너운반선 청보여운68987 충돌사건	2017-11-01
15204	원양어선 신유한호 해양오염사건	2017-05-01
15205	원양어선 카피딴 랠리에브(KAPITAN FALEYEV) 해양오염사건	2017-11-01
15206	어선 카피탄마슬로넥츠-어획물운반선 챌시8 충돌사건	2019-07-01
15207	원양어선 세종 기관손상사건	2020-12-01

2. 충돌사건 추출

```
arima_df_month = pd.DataFrame(df[df['해양사고종류(통계용)'].str.contains('충돌')]['해양사고발생시간(월별)'].value_counts()).sort_index()
arima_df_month.head()
```

	해양사고발생시간(월별)
2016-01-01	38
2016-02-01	32
2016-03-01	26
2016-04-01	26
2016-05-01	22

3. 현황파악 (시각화)



[그림 2-9] 시계열분석 준비

[데이터 검증]

1. 정상성 검정

```
> adf.test(df)

        Augmented Dickey-Fuller Test

data:  df
Dickey-Fuller = -3.8283, Lag order = 3, p-value = 0.02321
alternative hypothesis: stationary
```

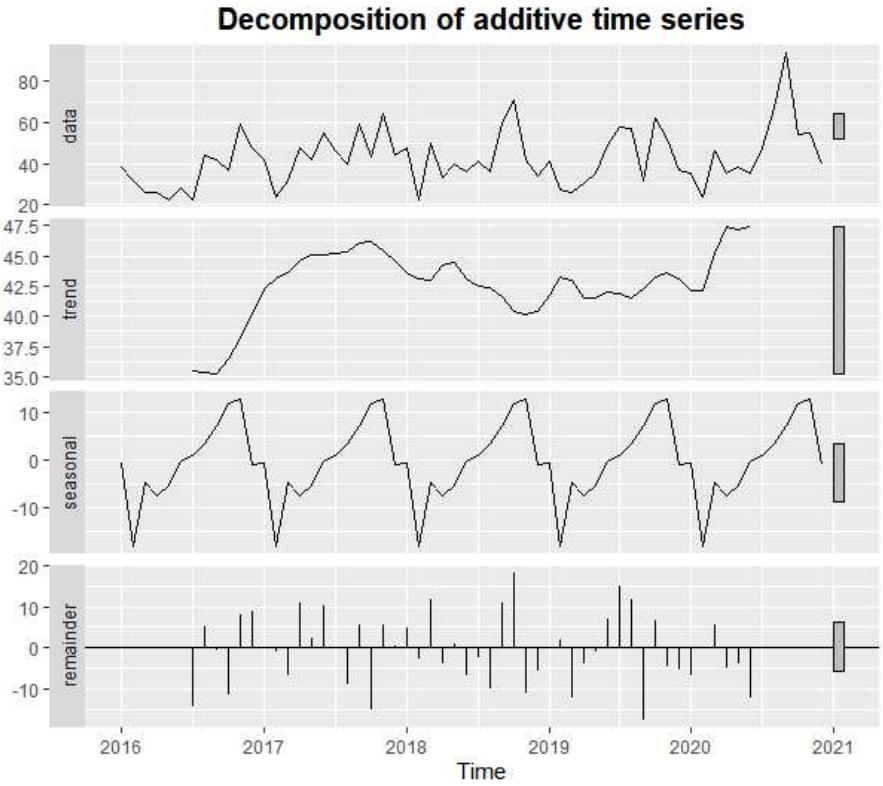
: p-value값이 0.05보다 작아 정상성을 만족하며 차분은 불필요하다.

2. ndiffs

```
> ndiffs(df)
[1] 0
```

: ndiffs함수 결과에서도 데이터의 차분을 추천하지는 않는다.

3. 분해시계열

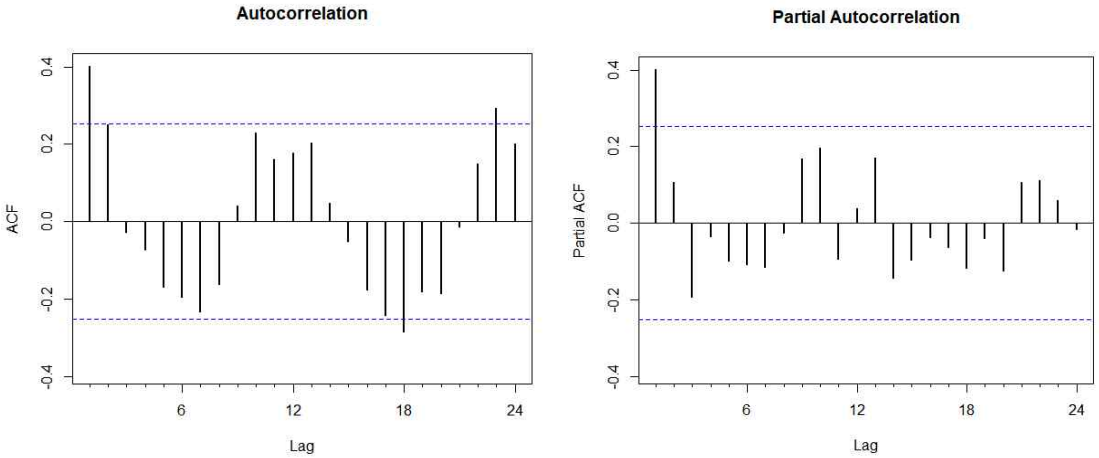


: 데이터 초반과 후반을 제외하면 큰 추세는 없는 것으로 판단되며, 약간의 계절성이 존재한다고 보인다.

[그림 2-10] 데이터 검증

[예측모델 생성]

1. ACF / PACF 확인



: ACF와 PACF 모두 시차1 이후에 0에 수렴한다.
따라서 AR(1)/ARMA(1,0), MA(1)/ARMA(0,1), ARMA(1,1) 등의 모델을 활용할 수 있다.

2. 최적 (p,d,q)도출

```
> auto.arima(df)
Series: df
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1      mean
      0.3967  42.4739
s.e.  0.1169   2.6579

sigma^2 estimated as 163.1:  log likelihood=-237.03
AIC=480.06   AICc=480.49   BIC=486.34
```

: auto.arima함수에서 본 데이터는 계절성이 없으며, AR(1)모델 활용을 추천하고 있다.

(*해양사고가 겨울철에 주로 발생하는 듯 한 계절성은 조업의 휴어기로 인한 것이며, 이는 분해시계열 그래프에서 아주 약한 계절성으로 확인된 바, ARIMA모델에서도 본 데이터셋은 계절성이 없는 데이터로 판단하였다.)

```
> arima(df, order = c(1,0,0))

Call:
arima(x = df, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.3967    42.4739
s.e.  0.1169     2.6579

sigma^2 estimated as 157.6:  log likelihood = -237.03,  aic = 480.06
> arima(df, order = c(0,0,1))

Call:
arima(x = df, order = c(0, 0, 1))

Coefficients:
         ma1  intercept
      0.2859    42.5081
s.e.  0.0980     2.1327

sigma^2 estimated as 166.2:  log likelihood = -238.58,  aic = 483.17
> arima(df, order = c(1,0,1))

Call:
arima(x = df, order = c(1, 0, 1))

Coefficients:
         ar1      ma1  intercept
      0.5080  -0.1301    42.4647
s.e.  0.2059   0.2208     2.8172

sigma^2 estimated as 156.7:  log likelihood = -236.86,  aic = 481.73
```

: MA(1)/ARMA(0,1), ARMA(1,1) 모델과의 직접 비교 결과 AR(1) 모델에서 AIC가 480.06으로 가장 낮은 것을 확인 할 수 있다.

3. 모델 생성

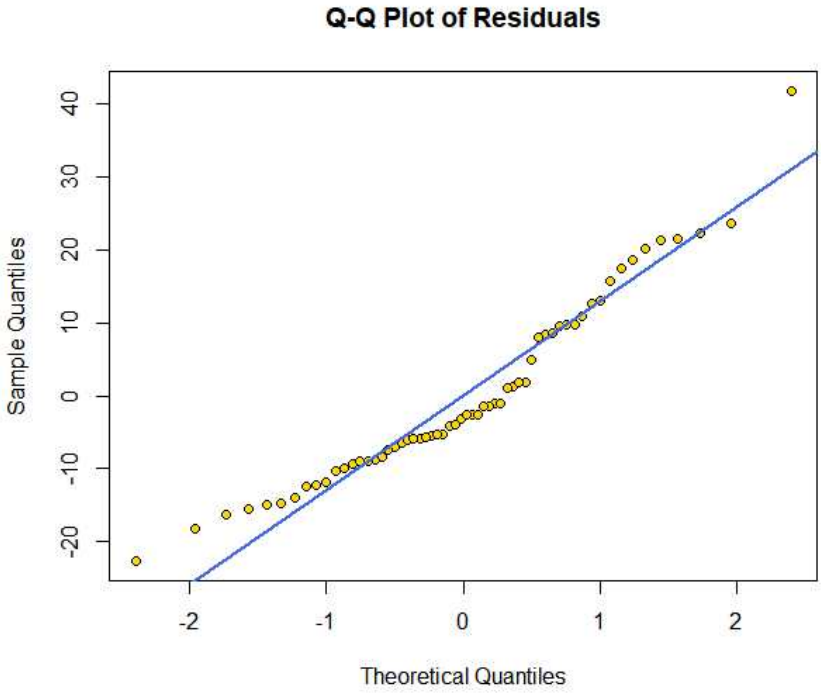
```
> df.arima <- arima(df, order = c(1, 0, 0))
> df.arima

Call:
arima(x = df, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.3967    42.4739
s.e.  0.1169     2.6579

sigma^2 estimated as 157.6:  log likelihood = -237.03,  aic = 480.06
```

: 가장 좋은 성능을 보인 AR(1)모델 (order=(1,0,0))으로 모델 생성

[그림 2-11] 예측모델 생성

[모델 평가]

1. Q-Q plot



: qqline(정규분포의 1Q와 3Q를 지나는 직선)을 크게 벗어나지 않아 데이터가 정규성 가정을 만족한다.

2. Box-Ljung test

```
> Box.test(df.arima$residuals, type = 'Ljung-Box')

        Box-Ljung test

data:  df.arima$residuals
X-squared = 0.080414, df = 1, p-value = 0.7767
```

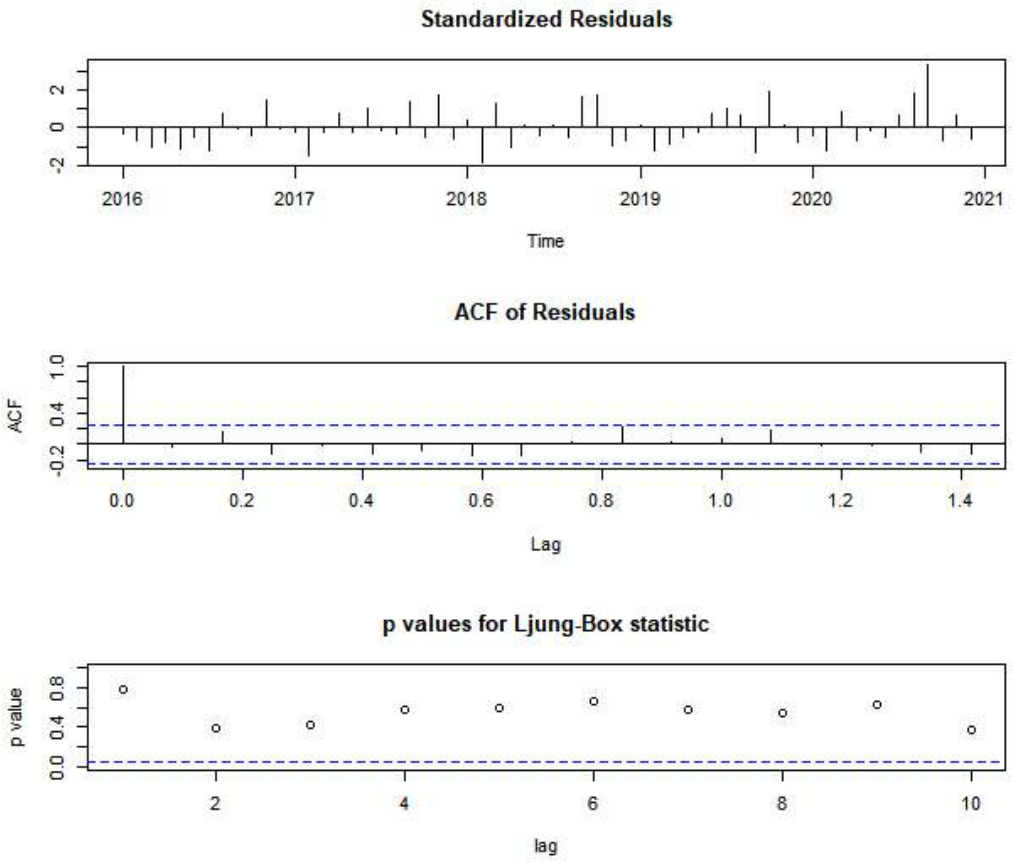
: 모형의 잔차를 이용하여 카이제곱검정 방법으로 검증한 Box-Ljung test 결과 p-value 0.05 이상이므로 자기상관(잔차간의 상관관계) 0이라는 귀무가설을 기각하지 못하고, 이는 데이터가 독립적으로 분포되어있음을 보여준다.

3. Accuracy

```
> accuracy(df.arima)
                     ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 0.03564687 12.55518 10.08948 -8.546437 25.45676 0.8290802 -0.03571249
> accuracy(arima(df, order = c(0,0,1)))
                     ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set 0.01157328 12.89358 10.13828 -9.375181 26.07035 0.8330896 0.08779376
> accuracy(arima(df, order = c(1,0,1)))
                     ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 0.05282741 12.51949 9.994219 -8.427782 25.18413 0.821252 -0.008179984
```

: 선정한 예측모델이 MA(1), ARMA(1,1) 모델들에 비해 평균적으로 작은 오차를 가지고 있으며, 현재 모델이 가장 좋은 성능을 띄는 것을 확인 할 수 있다.

4. 잔차분석



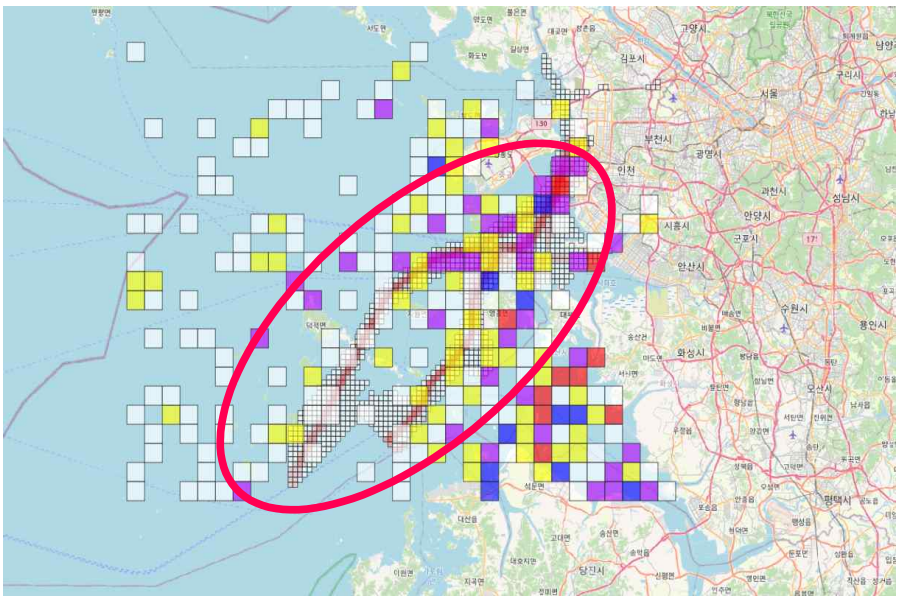
: ACF(자기상관함수)가 시차1 이상의 시점에서 모두 파란선(임계치) 안에 들어있으며, 이는 자기상관관계가 없고 데이터가 독립적임을 의미한다. 또한 p-value값이 모두 0 이상으로 분포하여 본 모델은 매우 양호한, 적합한 모델이라는 것을 알 수 있다.

[그림 2-12] 모델평가

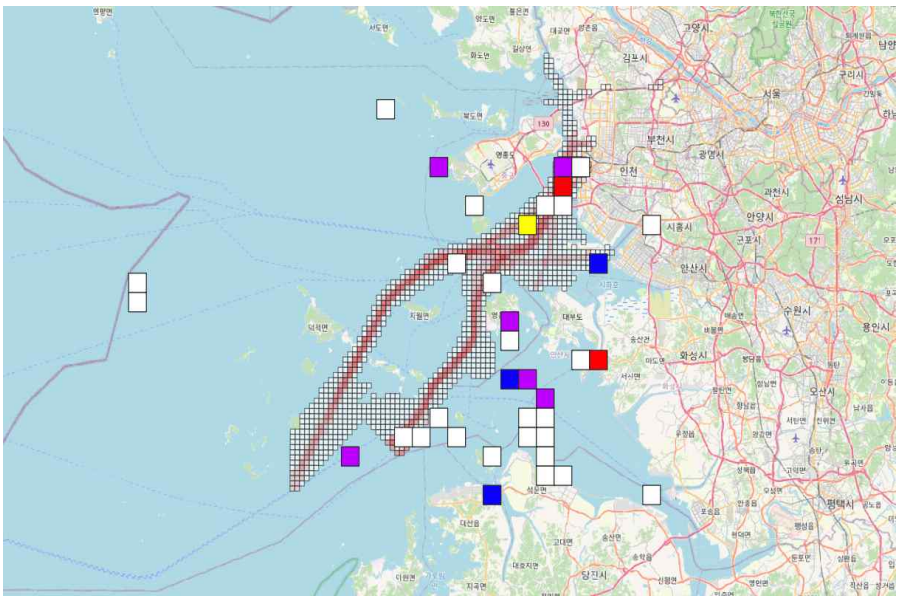
3. 분석 결과

3-1. 수행 결과

□ GIS 공간분석



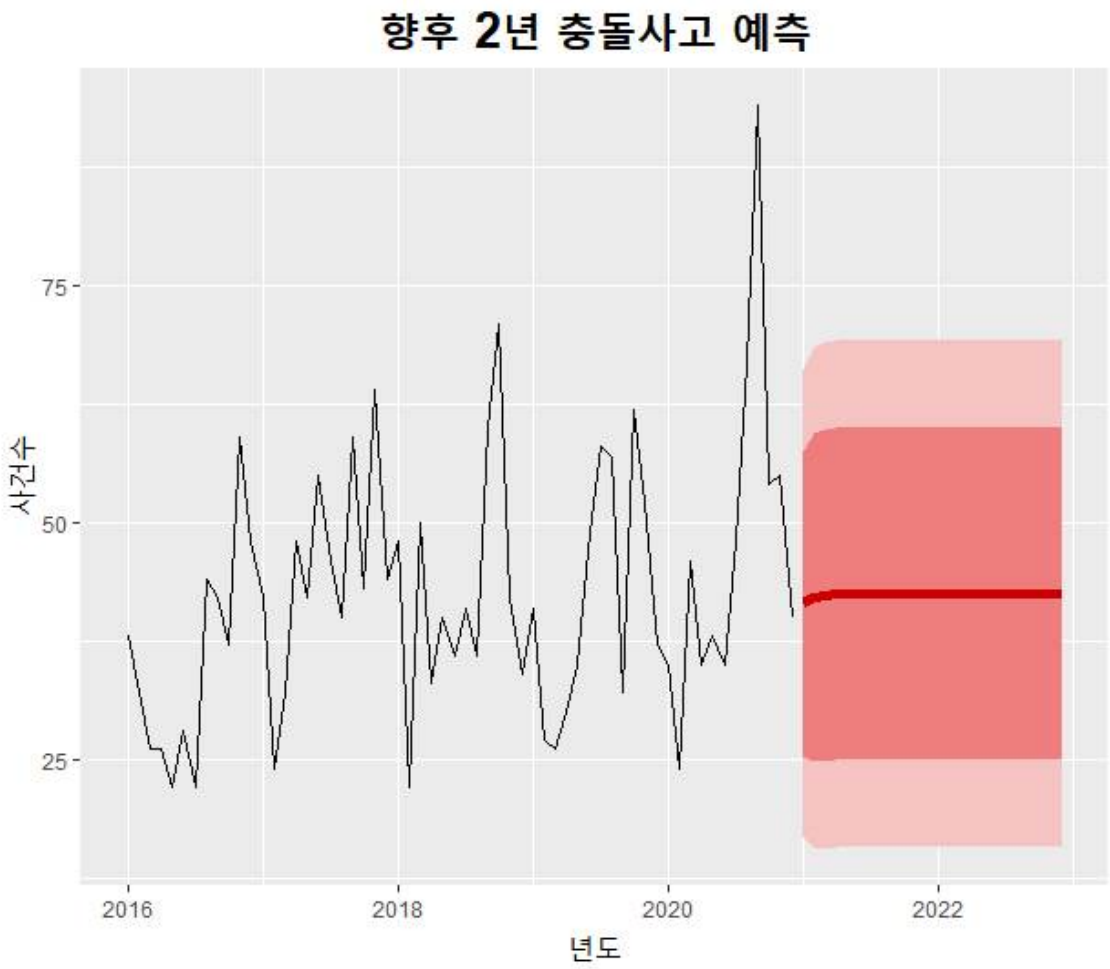
[그림 3-1] 항적도내 어선 분포현황



[그림 3-1] 항적도 내 충돌사고 분포현황

: 도출해낸 주요 위험 어선사고지역과 충돌 위험지역을 대형선박(화물선, 유조선, 여객선 등) 주요 항적도 위에서 관찰한 결과 실제로 주요 항로 내에서 어선이 다수 분포하며 이는 충돌사건의 주요 원인이 될 수 있음을 알 수 있다.

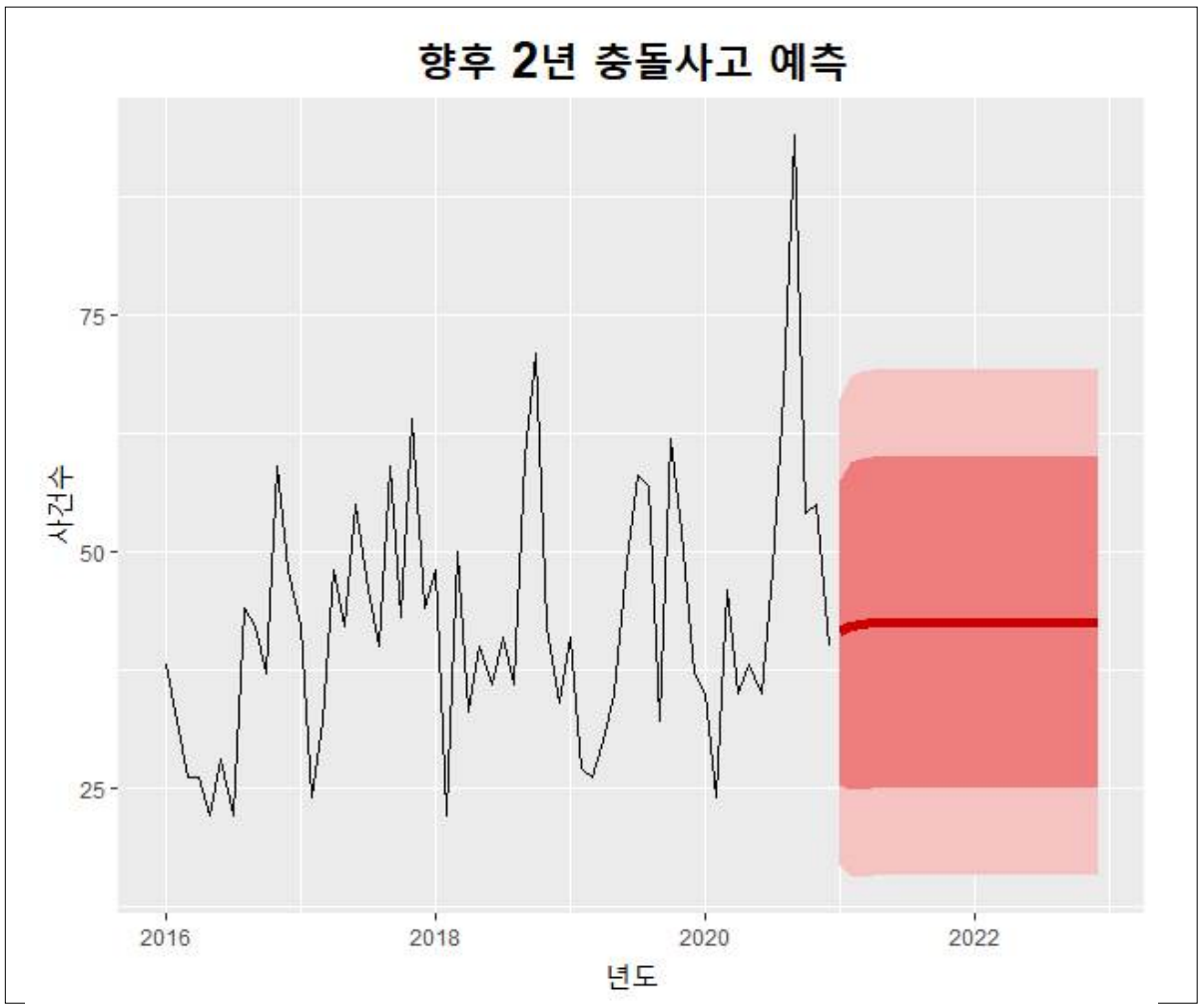
□ 시계열 분석



[그림 3-2] 향후 2년간의 충돌사고 예측 그래프

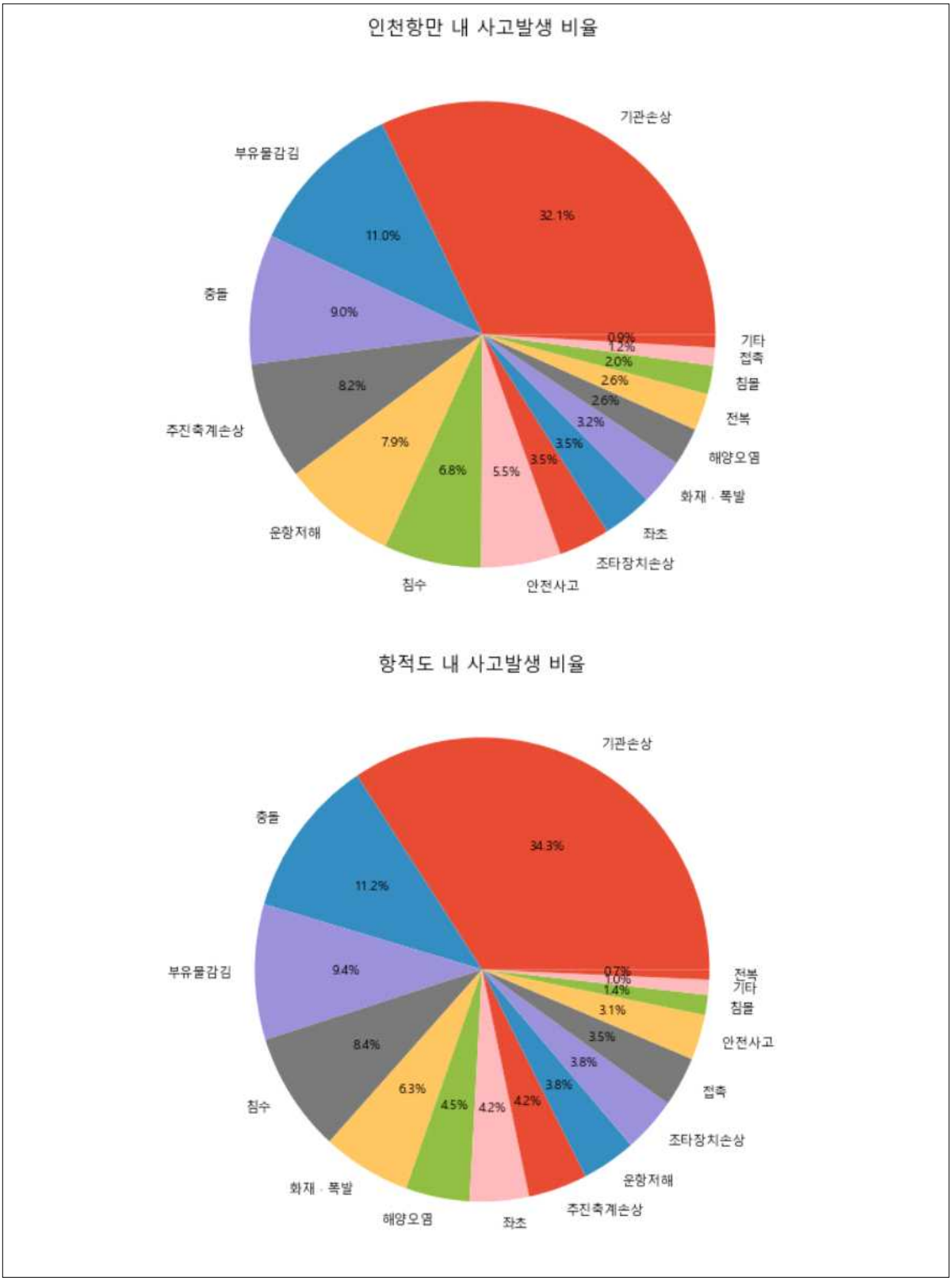
: 분홍색과 빨강색 바탕은 각각 80%와 95%의 신뢰구간에서의 2022년도까지의 해양사고 발생 예측을 나타내며, 빨강 실선은 예측값들의 평균을 나타낸다. 따라서 평균적인 해양사고 발생 숫자는 약간의 증가가 예상지만 전체적 추세의 변동은 크지 않을 것으로 해석된다.

4. 결론



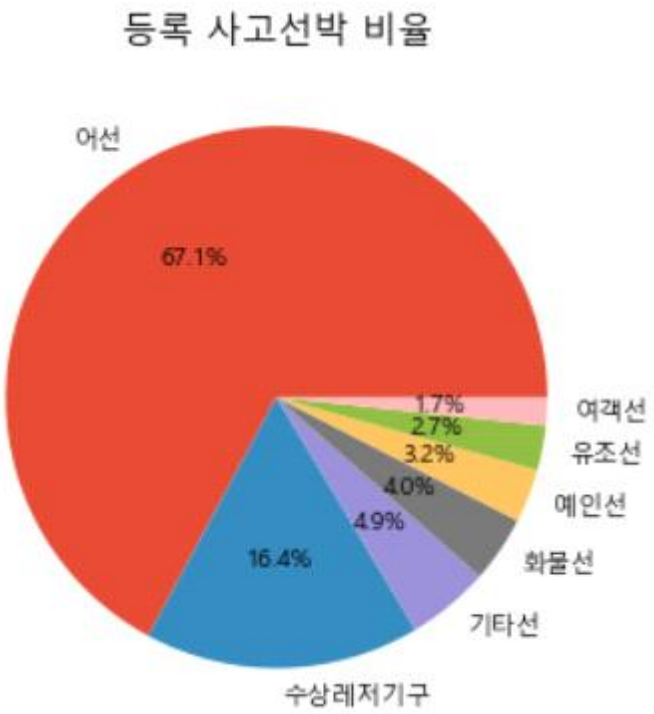
[그림 4-1] 향후 2년간의 충돌사고 예측 그래프

2022년에도 해양하고의 추세에는 큰 변동 없이 예년과 비슷한 수준에서 발생할 것으로 예상된다. 이러한 현상은 현재의 해양사고 방지 대책이 효과적으로 활용되지 못하다는 것을 의미하여 보다 구체적이고 현실적으로 추세를 낮출 수 있는 방안에 대해 탐구하였다.



[그림 4-2] 인천항만 전체 사고비율 vs 항적도 내 사고비율

항적도 내에서의 충돌사고 발생률은 11%로 항만 전 범위에 비해 2%가량 높은 것으로 나타났다. 이는 충돌사고가 선박 주요 항로 내에 집중 분포하고 있는 것을 의미한다.



[그림 4-3] 등록 사고선박 비율

본 프로젝트의 안전관리 촉구 자료로써의 활용 근거로 다음의 3가지를 제시 할 수 있다.

- 항적도상에서의 충돌사건 발생 위험성이 높다.
- 항적도상에서 어선의 해양사고 발생이 집중되어있다.
- 등록 사고선박으로 어선이 과반수이상을 차지하고 있다.

따라서 선제적으로 어민들의 충돌사고에 대한 경각심을 일깨워 충돌사고에 관한 안전관리가 보다 효과적으로 실천된다면 전반적인 충돌사고의 발생 추세를 크게 감소시킬 수 있을 것으로 기대된다.

□ 프로젝트 활용방안

- 주 항로상의 충돌 위험지역에 위치할 시 경고등(사이렌)을 울림으로써 데이터로는 나타나지 않는 충돌사건의 주된 인적 요인인 '졸음운항'을 방지해 충돌사고 및 기타 인적 해양사고를 예방한다.

- 레이더에 현재 프로젝트에서 제시한 위험지역을 표기하여 어민들의 경각심을 고조시킬 수 있다.

□ 한계점

개인정보보호법에 의한 일부 데이터 접근 불가로 해양사고 당시의 기상, 인적사항, 선박정보 등의 데이터가 부재하였다. 따라서 회귀모델로서 어선이 충돌사고에 미치는 영향도에 대한 분석은 불가능하였고, 수치화를 시키는 데에는 한계가 존재했다.