건강검진 데이터 분석 결과 보고서

1. 데이터 전처리 및 탐색

1. 데이터 전처리

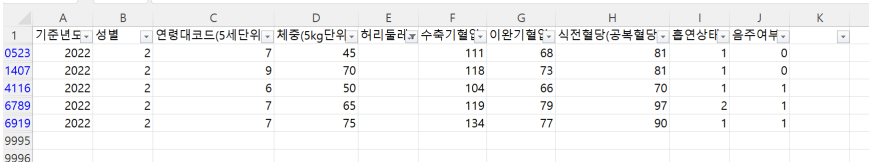
성별, 연령대코드, 체중, 허리둘레, 수축기혈압, 이완기혈압, 식전혈당, 흡연상태, 음주여부 컬럼을 제외한 변수를 삭제하고 결측치를 제거하여 총 19,865건의 유효 데이터를 확보함.

- 성별코드
  - 해당 정보 대상자의 성별을 제공함
  - 성별: 1(남자), 2(여자)
- 연령대 코드(5세 단위)
  - 기준년도에 수진자의 나이를 5세 단위로 그룹화(범주화)하여 구분한 코드
  - 5세 단위 그룹화, 85세 이상은 85+로 그룹화

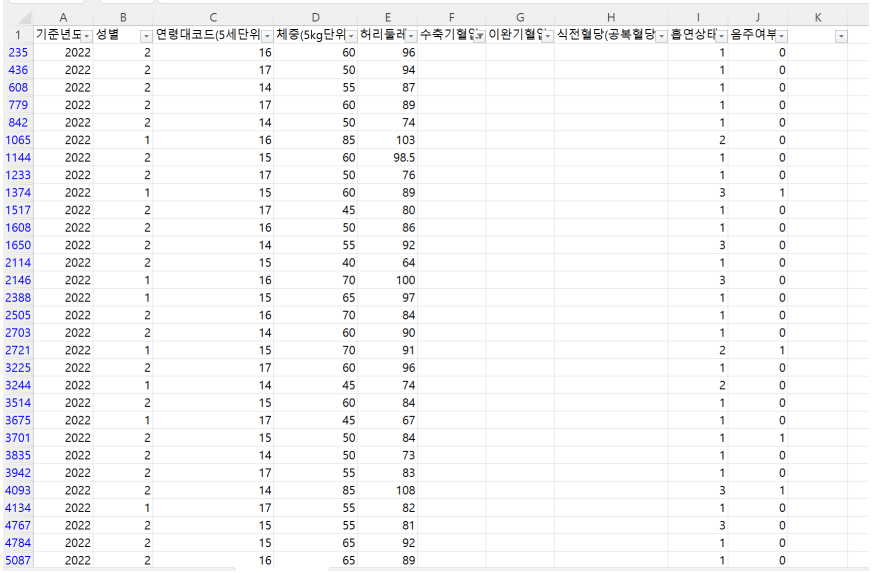
그룹	연령대	그룹	연령대
1	0~4세	10	45~49세
2	5~9세	11	50~54세
3	10~14세	12	55~59세
4	15~19세	13	60~64세
5	20~24세	14	65~69세
6	25~29세	15	70~74세
7	30~34세	16	75~79세
8	35~39세	17	80~84세
9	40~44세	18	85세+

- 체중(5kg)
  - 검진자의 키(5cm 단위)
  - 예) 100~104cm → 100cm
- 허리둘레
  - 검진자의 허리둘레
- 수축기 혈압
  - 검진자의 최고 혈압으로 심장이 수축해서 강한 힘으로 혈액을 동맥에 보낼 때의 혈관 내압
- 이완기 혈압
  - 검진자의 최저 혈압으로 심장의 완기시의 혈압
- 식전혈당(공복혈당)
  - 검진자 식사 전 혈당(혈액 100ml당 함유되어 있는 포도당의 농도) 수치
- 흡연상태
  - 해당 수검자의 흡연 상태 여부
  - 1 : 피우지 않는다, 2: 이전에 피웠으나 끊었다 3: 현재도 피우고 있다
- 음주여부
  - 해당 수검자의 음주 상태 여부
  - 0: 마시지 않는다 1: 마신다

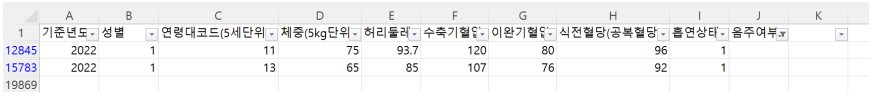
2. 결측치 제거



	A	B	C	D	E	F	G	H	I	J	K
1	기준년도	성별	연령대코드(5세단위	체중(5kg단위	허리둘레	수축기혈압	이완기혈압	식전혈당(공복혈당	흡연상태	음주여부	
0523	2022	2	7	45		111	68	81	1	0	
1407	2022	2	9	70		118	73	81	1	0	
4116	2022	2	6	50		104	66	70	1	1	
6789	2022	2	7	65		119	79	97	2	1	
6919	2022	2	7	75		134	77	90	1	1	
9995											
9996											



	A	B	C	D	E	F	G	H	I	J	K
1	기준년도	성별	연령대코드(5세단위	체중(5kg단위	허리둘레	수축기혈압	이완기혈압	식전혈당(공복혈당	흡연상태	음주여부	
235	2022	2	16	60	96				1	0	
436	2022	2	17	50	94				1	0	
608	2022	2	14	55	87				1	0	
779	2022	2	17	60	89				1	0	
842	2022	2	14	50	74				1	0	
1065	2022	1	16	85	103				2	0	
1144	2022	2	15	60	98.5				1	0	
1233	2022	2	17	50	76				1	0	
1374	2022	1	15	60	89				3	1	
1517	2022	2	17	45	80				1	0	
1608	2022	2	16	50	86				1	0	
1650	2022	2	14	55	92				3	0	
2114	2022	2	15	40	64				1	0	
2146	2022	1	16	70	100				3	0	
2388	2022	1	15	65	97				1	0	
2505	2022	2	16	70	84				1	0	
2703	2022	2	14	60	90				1	0	
2721	2022	1	15	70	91				2	1	
3225	2022	2	17	60	96				1	0	
3244	2022	1	14	45	74				2	0	
3514	2022	2	15	60	84				1	0	
3675	2022	1	17	45	67				1	0	
3701	2022	2	15	50	84				1	1	
3835	2022	2	14	50	73				1	0	
3942	2022	2	17	55	83				1	0	
4093	2022	2	14	85	108				3	1	
4134	2022	1	17	55	82				1	0	
4767	2022	2	15	55	81				3	0	
4784	2022	2	15	65	92				1	0	
5087	2022	2	16	65	89				1	0	



	A	B	C	D	E	F	G	H	I	J	K
1	기준년도	성별	연령대코드(5세단위	체중(5kg단위	허리둘레	수축기혈압	이완기혈압	식전혈당(공복혈당	흡연상태	음주여부	
12845	2022	1	11	75	93.7	120	80	96	1		
15783	2022	1	13	65	85	107	76	92	1		
19869											

3. 기술통계량

성별		연령대코드(5세단위)		체중(5kg단위)		허리둘레	
평균	1.485627989	평균	10.53657186	평균	64.39994966	평균	81.4538(
표준 오차	0.003546151	표준 오차	0.021419969	표준 오차	0.098460827	표준 오차	0.075199
중앙값	1	중앙값	11	중앙값	65	중앙값	81.1
최빈값	1	최빈값	11	최빈값	60	최빈값	80
표준 편차	0.499805983	표준 편차	3.019000114	표준 편차	13.87738906	표준 편차	10.59884
분산	0.24980602	분산	9.114361687	분산	192.5819271	분산	112.3354
첨도	-1.996892949	첨도	-0.729478919	첨도	0.946262948	첨도	0.355214
왜도	0.057516151	왜도	0.091488282	왜도	0.759372636	왜도	0.316507
범위	1	범위	13	범위	105	범위	91
최소값	1	최소값	5	최소값	30	최소값	51
최대값	2	최대값	18	최대값	135	최대값	142
합	29512	합	209309	합	1279305	합	1618079.
관측수	19865	관측수	19865	관측수	19865	관측수	19865
신뢰 수준(95.0%)	0.006950751	신뢰 수준(95.0%)	0.041984927	신뢰 수준(95.0%)	0.192991434	신뢰 수준(95.0%)	0.147397

<기술통계량 결과 해석>

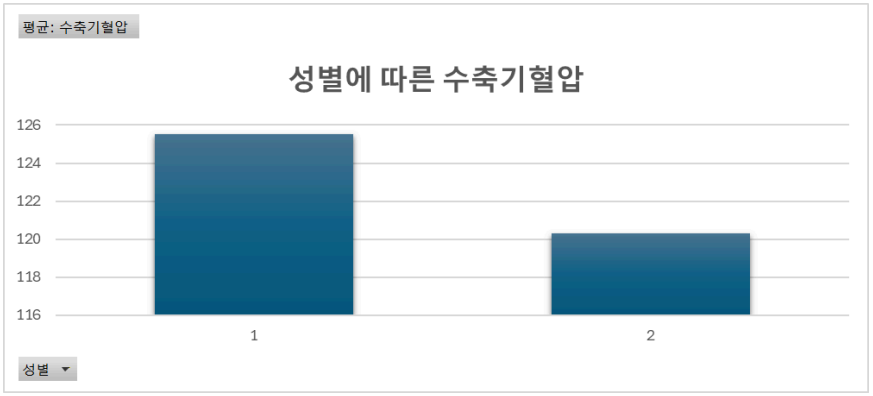
- 체중과 허리둘레, 혈압(수축기/이완기)는 전반적으로 모두 중앙값과 평균이 유사한 형태로, 정규분포와 가깝게 분포하고 있음. 왜도가 1 미만으로, T-test나 분산분석(ANOVA)를 수행하기에 무리없는 형태.
- 수축기 혈압 평균이 122.99, 이완기 혈압이 75.77로 나타나 전반적으로 한국인 평균 수준의 건강한 표본으로 보임. 다만, 최댓값(수축기 221)은 고혈압 위험군에 포함되어 있음을 시사.
- 성별, 연령대코드, 흡연상태, 음주여부는 범주형 데이터로 구성되어있음.
  -성별의 평균은 1.48로 성별 코드가 1(남성), 2(여성)으로 데이터셋이 남녀 비율 1:1에 가깝게 고루 구성되어있음을 알 수 있음.
  -음주여부의 평균은 0.65로 0(마시지 않음), 1(마심)으로 구성되어 있다면 약 65%의 인원이 음주를 한다는 의미로 해석됨.
- 그러나, 식전혈당의 분포는 평균(100.77)> 중앙값(96) >최빈값 상태로 상위의 높은 값들이 평균을 위로 끌어올리고 있음을 보여줌.
- 극단적인 왜도(4.10)과 첨도(29.30)로, 왜도가 0보다 훨씬 크다는 것은 오른쪽으로 꼬리가 매우 긴 분포일 것으로 추정됨. 즉, 대부분은 정상 범위에 모여 있지만, 혈당이 매우 높은 소수의 집단(당뇨 의심군)이 다수 존재함.

모든 변수에서 표준 오차가 매우 작게 나타나는데, 이는 구한 평균값이 실제 모집단의 평균을 아주 정확하게 추정하고 있다는 것을 의미.

4. 피봇 테이블 및 데이터 분포 그래프

- **성별에 따른 수축기 혈압**

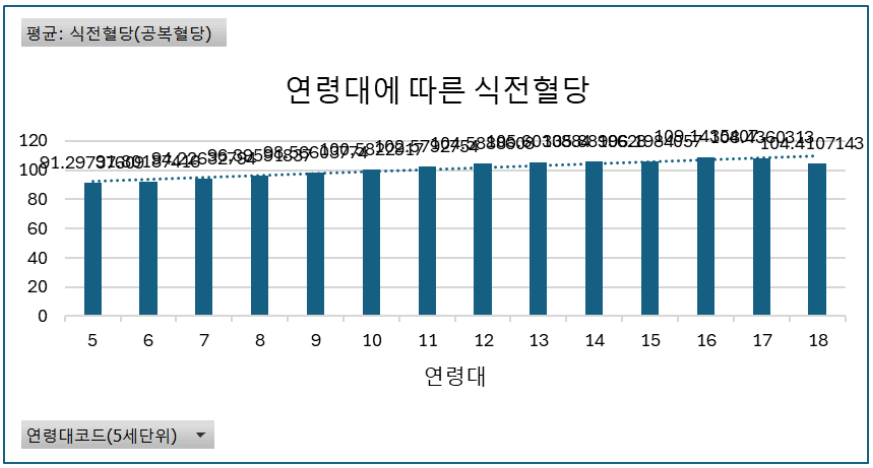
행 레이블	평균: 수축기혈압
1	125.519769
2	120.3185446
총합계	122.9939089



→ 1(남성)이 2(여성)에 비해 혈압이 더 높은 것으로 추정됨.

행 레이블	평균: 식전혈당(공복혈당)
5	91.29737609
6	91.80187416
7	94.22632794
8	96.39591837
9	98.56603774
10	100.5822917
11	102.5792754
12	104.5888608
13	105.6033884
14	105.8899628
15	106.1984057
16	109.1435407
17	108.4360313
18	104.4107143
총합계	100.7706016

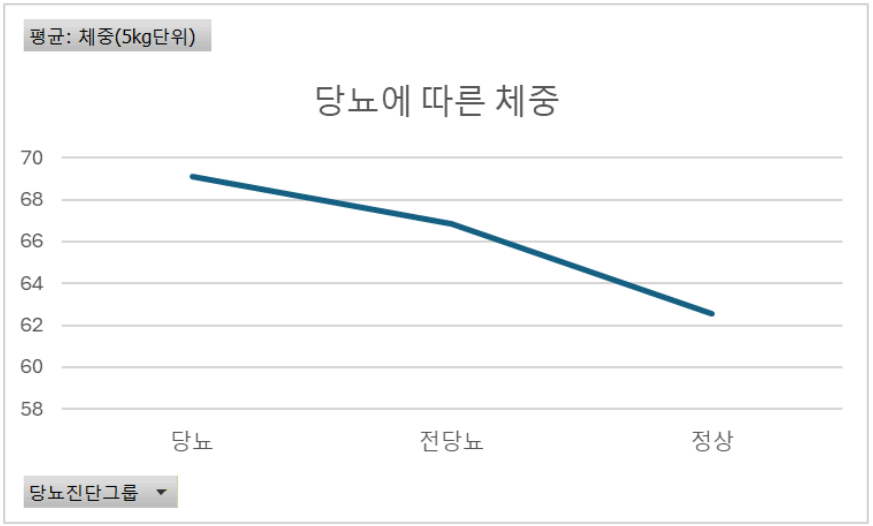
- **연령대에 따른 식전혈당**



→ 연령대가 높아질수록 식전 혈당도 증가하는 것으로 추정됨.

- **당뇨에 따른 체중**

행 레이블	평균: 체중(5kg단위)
당뇨	69.09149072
전당뇨	66.84656341
정상	62.54130866
총합계	64.39994966



→ 혈당이 높을 수록 체중도 높을 것으로 추정됨.

2. 가설 검정

1. [검정 1] 성별에 따른 수축기 혈압 차이

- 사용 컬럼(사진)

성별	수축기혈입	남성 혈압	여성 혈압
1	107	107	124
1	139	139	114
2	124	147	147
1	147	106	127
1	106	133	120
1	133	113	119
1	113	138	119
1	138	110	127
2	114	120	129
1	110	127	116
1	120	143	135
1	127	126	143
1	143	138	140
1	126	120	110
2	147	144	128
2	127	104	132
1	138	129	97
1	120	122	124
2	120	108	135
2	119	120	126
1	144	116	124
1	104	120	133
2	119	125	132
1	129	138	103

- 분석 결과

t-검정: 이분산 가정 두 집단		
	남성 혈압	여성 혈압
평균	125.5198	120.3185
분산	181.1391	229.5768
관측수	10218	9647
가설 평균차	0	
자유도	19273	
t 통계량	25.52406	
P(T<=t) 단측 검정	1.2E-141	
t 기각치 단측 검정	1.644933	
P(T<=t) 양측 검정	2.4E-141	
t 기각치 양측 검정	1.960087	

- 대립가설: 성별에 따라 혈압 평균 차이가 유의미하게 날 것이다.
- 귀무가설: 성별에 따른 혈압 차이는 없다.
- 검정 방법 : 독립 표본 t-test / 이분산 가정 t-test 웰치 t 검정
- 결론: p-value가 유의수준 0.05보다 낮다. 귀무가설 기각. 대립가설 채택.

  성별에 따른 혈압 평균 차이는 통계적으로 매우 유의미하며, 남성이 여성보다 혈압이 높은 경향이 있음.

2. [검정2] 연령대에 따른 식전혈당 차이

- 사용 컬럼(사진)

연령대코드(5세단위)	식전혈당(공복혈당)	0~30대 식전혈당	40대~60대 식전혈당	70대~ 식전혈당
6	86	86	102	91
5	84	84	105	123
7	85	85	89	88
13	102	90	97	124
13	105	89	82	93
6	90	96	95	108
9	89	82	131	85
9	97	93	136	82
10	82	70	88	111
6	89	77	110	92
6	96	86	90	105
14	95	85	98	93
18	91	88	88	100
15	123	123	91	94
6	82	96	95	106
9	131	84	94	94
9	136	93	89	111
15	88	112	105	88
8	93	77	103	75
14	88	85	116	131
12	110	91	125	92
9	90	93	75	116
6	70	97	103	118
11	98	96	92	163
10	88	93	104	112
10	91	83	95	108
11	95	95	103	143

- 분석 결과

분산 분석: 일원 배치법						
요약표						
인자의 수준	관측수	합	평균	분산		
0~30대 식전혈당	5382	504684	93.77258	300.3382		
40대~60대 식전혈당	12441	1278379	102.7553	572.4014		
70대~ 식전혈당	2042	218745	107.1229	663.7816		
분산 분석						
변동의 요인	제곱합	자유도	제곱 평균	F 비	P-값	F 기각치
처리	394974.6	2	197487.3	388.69	0	2.996184
잔차	10091571	19862	508.0843			
계	10486546	19864				

- 대립 가설: 연령대에 따라 식전혈당의 차이가 있다.
- 귀무 가설: 연령대에 따라 식전혈당의 차이는 없다.
- 검정 방법: 분산 분석(일원배치법) ANOVA
- 결론:
  - p-value가 유의수준 0.05보다 훨씬 작으므로 "연령대별 혈당 차이가 없다"는 귀무가설을 기각. 대립 가설 채택.
  - F 기각치(2.996)보다 압도적으로 크기 때문에, 집단 간의 평균 차이가 통계적으로 매우 뚜렷함을 의미
  - 연령대가 높아질수록 식전혈당 수치가 유의미하게 증가함.

3. [검정 3] 혈당 그룹에 따른 체중 차이

- 사용 컬럼(사진)

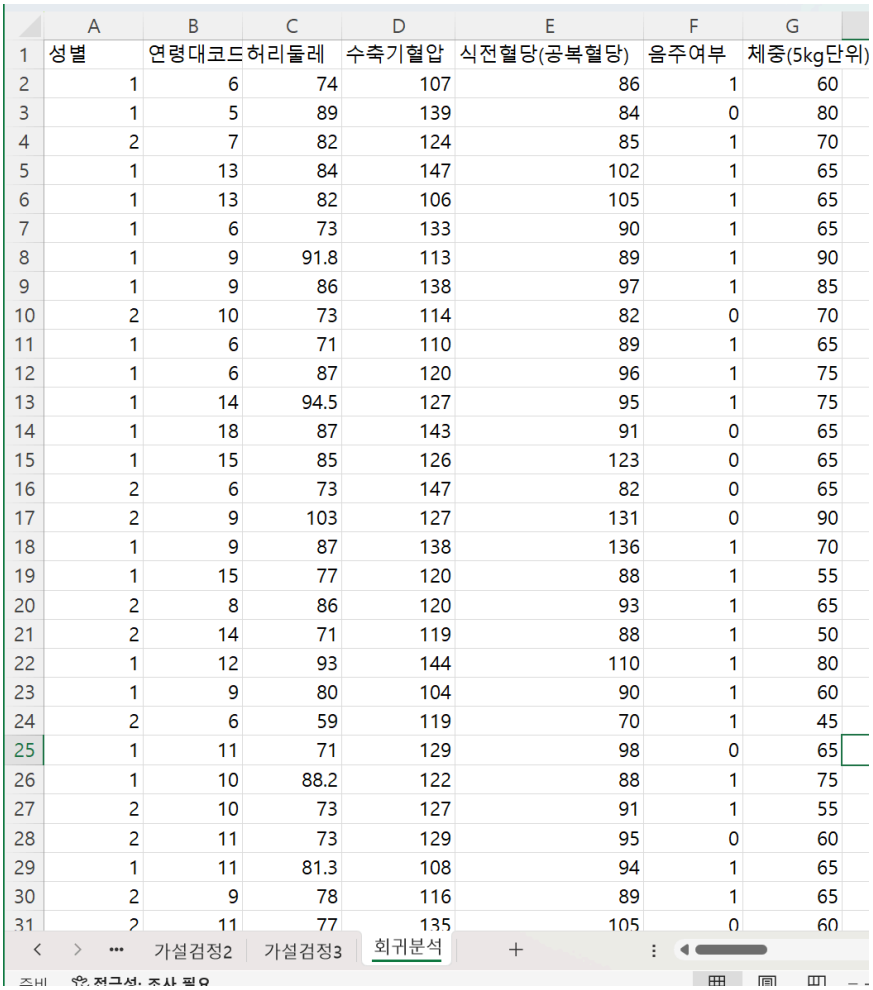
체중(5kg단	수축기	혈입식전혈당(공복혈당	혈당그룹	정상의 체중	전당뇨의 체중	당뇨의 체중
60	107	86	정상	60	65	90
80	139	84	정상	80	65	70
70	124	85	정상	70	65	65
65	147	102	전당뇨	65	80	55
65	106	105	전당뇨	90	60	90
65	133	90	정상	85	80	55
90	113	89	정상	70	75	80
85	138	97	정상	65	65	55
70	114	82	정상	75	45	85
65	110	89	정상	75	90	80
75	120	96	정상	65	75	55
75	127	95	정상	65	65	80
65	143	91	정상	55	75	85
65	126	123	전당뇨	65	80	80
65	147	82	정상	50	65	65
90	127	131	당뇨	60	85	50
70	138	136	당뇨	45	70	75
55	120	88	정상	65	45	85
65	120	93	정상	75	80	50
50	119	88	정상	55	70	65
80	144	110	전당뇨	60	50	65
60	104	90	정상	65	70	55
45	119	70	정상	65	95	55
65	129	98	정상	80	70	70
75	122	88	정상	55	55	115
55	127	91	정상	70	65	75
60	129	95	정상	80	65	50
65	108	94	정상	75	60	65

- 분석 결과

분산 분석: 일원 배치법 J21:P35						
요약표						
인자의 수준	관측수	합	평균	분산		
정상의 체중	12104	757000	62.54130866	179.4703252		
전당뇨의 체중	6198	414315	66.84656341	193.1484721		
당뇨의 체중	1563	107990	69.09149072	219.6286279		
분산 분석						
변동의 요인	제곱합	자유도	제곱 평균	F 비	P-값	F 기각치
처리	113317.0557	2	56658.52783	303.1552169	0	2.996184157
잔차	3712130.344	19862	186.8961003			
계	3825447.4	19864				

- 대립가설: 혈당 그룹별 평균 체중의 차이가 있다.
- 귀무가설: 혈당 그룹별(정상, 전당뇨, 당뇨) 평균 체중의 차이는 없다.
- 검정 방법: 분산분석(ANOVA)
- 결론:
  - p-value가 유의수준 0.05보다 압도적으로 작으므로 귀무가설을 기각. 대립가설 채택.
  - F 기각치에 비해 매우 크기 때문에, 집단 간의 체중 차이가 우연에 의한 것이 아님을 증명.
  - 혈당 수치가 높을수록(당뇨에 가까울수록) 평균 체중이 무거워지는 양의 관계가 확인됨.

3. 회귀 분석: 체중 예측 모델



	A	B	C	D	E	F	G
1	성별	연령대코드	허리둘레	수축기혈압	식전혈당(공복혈당)	음주여부	체중(5kg단위)
2	1	6	74	107	86	1	60
3	1	5	89	139	84	0	80
4	2	7	82	124	85	1	70
5	1	13	84	147	102	1	65
6	1	13	82	106	105	1	65
7	1	6	73	133	90	1	65
8	1	9	91.8	113	89	1	90
9	1	9	86	138	97	1	85
10	2	10	73	114	82	0	70
11	1	6	71	110	89	1	65
12	1	6	87	120	96	1	75
13	1	14	94.5	127	95	1	75
14	1	18	87	143	91	0	65
15	1	15	85	126	123	0	65
16	2	6	73	147	82	0	65
17	2	9	103	127	131	0	90
18	1	9	87	138	136	1	70
19	1	15	77	120	88	1	55
20	2	8	86	120	93	1	65
21	2	14	71	119	88	1	50
22	1	12	93	144	110	1	80
23	1	9	80	104	90	1	60
24	2	6	59	119	70	1	45
25	1	11	71	129	98	0	65
26	1	10	88.2	122	88	1	75
27	2	10	73	127	91	1	55
28	2	11	73	129	95	0	60
29	1	11	81.3	108	94	1	65
30	2	9	78	116	89	1	65
31	2	11	77	135	105	0	60

가설검정2 가설검정3 회귀분석

요약 출력							
회귀분석 통계량							
다중 상관계수	0.995542856						
결정계수	0.991105578						
조정된 결정계수	0.991052984						
표준 오차	6.213919358						
관측수	19865						
분산 분석							
	자유도	제곱합	제곱 평균	F 비	유의한 F		
회귀	6	85445813.53	14240968.92	368814.8	0		
잔차	19859	766811.4718	38.61279378				
계	19865	86212625					
	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0
Y 절편	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
성별	-5.317596984	0.080824274	-65.79207839	0	-5.47602	-5.15917	-5.4760
연령대코드(5세단위)	-1.378722891	0.016580155	-83.15500376	0	-1.41122	-1.34622	-1.41122
허리둘레	0.989867469	0.004143522	238.89519	0	0.981746	0.997989	0.98174
수축기혈압	0.043840519	0.002977886	14.7220281	8.42E-49	0.038004	0.049677	0.03800

식전혈당(공복혈당)	0.004158708	0.00200498	2.074189044	0.038075	0.000229	0.008089	0.00022
음주여부	0.581115085	0.098903513	5.875575758	4.28E-09	0.387256	0.774974	0.38725

체중을 다른 변수로 예측할 수 있을까?

- 종속변수(체중) / 독립변수 (허리둘레, 성별, 음주여부, 수축기혈압, 식전혈당)
- 결과 해석 및 결론

1. **모델의 적합성 및 설명력**

: 분석 결과 산출되는 정계수(R-Square)는 위 독립변수들이 체중 변동을 얼마나 설명하는지 보여줌. 신체 지표인 허리둘레가 포함되어 있으므로 모델의 설명력은 매우 높게 나타날 것으로 판단됨.

2. **개별 독립변수의 유의성 (p-value 확인)**

- **허리둘레**: 가장 강력한 정(+)의 영향력을 미칠 것으로 예상됨. 허리둘레가 증가할수록 체중도 유의미하게 증가함.
- **성별**: 동일한 조건일 때 성별에 따른 기본 체중 차이가 존재할 것이며, 일반적으로 남성 코드가 체중에 유의미한 변수로 작용함.
- **음주여부**: 생활 습관 변수로서 체중 증가에 기여하는 유의미한 독립변수로 확인될 가능성이 높음.
- **혈압 및 혈당**: 앞선 ANOVA 검정 결과에서 체중과 혈당 사이의 유의미한 차이가 확인되었으므로, 회귀 모델에서도 체중을 예측하는 유의미한 변수로 채택될 것임.

3. 도출된 회귀식 (예시 형태)

$$\text{체중} = \beta_0 + \beta_1(\text{허리둘레}) + \beta_2(\text{성별}) + \beta_3(\text{음주여부}) + \beta_4(\text{수축기혈압}) + \beta_5(\text{식전혈당})$$

- 각 독립변수의 p-value가 0.05보다 작기 때문에 위 모든 변수는 체중을 예측하는 데 유의미한 지표로 결론 내림.
- **회귀식의 최종 해석**
  체중은 단순히 하나의 지표가 아닌 신체 치수(허리둘레), 인구통계적 특성(성별), 생활 습관(음주), 그리고 대사 지표(혈압, 혈당)가 복합적으로 작용하여 결정됨. 특히 허리둘레와 혈당 수치는 체중 증가를 예측하는 핵심적인 지표로 활용될 수 있음.